I

INDIGENOUS INTERVIEWERS IN SURVEY RESEARCH

Chairman, ANTHONY PASCAL, Rand Corporation

Page

THE USE OF A NEIGHBORHOOD SURVEY TEAM IN THE EVALUATION OF A YOUTH EMPLOYMENT PROJECT

P. O. Katsky and R. F. Wickland, The RAND Corporation

Techniques for maximizing the quality and quantity of interview response from "non-middle class" populations are in need of development. Past research has pointed out some of the sources of bias resulting from the use of interviewers of the same and higher class and caste status.

A description is given of a study using a team of indigenous interviewers designed to overcome some of the inadequacies of procedures used in the past to interview ghetto populations. The team was composed of individuals matched to respondents in such characteristics as age, sex, social class, residence, and past experience. Methods of team recruitment, training, and management are described with references to differences between the operation of this team and surveys using professional interviewers. A typology of team members and their differing contributions to the research effort is presented.

An evaluation of the success of the indigenous interviewers in producing quality interviews and in locating respondents is presented along with suggestions for new directions and refinements.

1. INTRODUCTION

This paper describes the use of an experimental technique for interviewing other than "middle" and "upper" class populations. This technique -- the creation of a Neighborhood Survey Team (NST) staffed by individuals similar in background to those being interviewed -- was developed and tested to interview black and brown youth who had participated in an anti-poverty employment training program.

Although using the NST seemed to us to be a good idea from several standpoints, our main research task was to evaluate a youth training program and not to test the NST idea in any formal or systematic way. In light of this fact, this paper does not contain a statistically elegant treatment of validity resulting from use of indigenous interviewers (a subject of interest to survey statisticians and amenable to statistical analysis), but rather reports on some tentative lessons learned as a result of employing this technique. Although the experience with both the black and the brown team was substantially similar, only the Mexican-American (brown) team, which was in existence longer and completed more interviews, will be discussed here.[1]

The paper is divided into the following sections:

Interviewing the Ghetto Resident: Theory
Description of the NST
Conclusions

2. INTERVIEWING THE GHETTO RESIDENT: THEORY

Some of the difficulties encountered by recent survey efforts in ghetto areas have included poor location rates for respondents, poor quality of response, and high community antagonism to interviewing procedures. Hypotheses have been advanced that individuals not socialized to middle class values may not consider the interviewing situation one deserving of full attention and seriousness.

In order to compensate for these factors tending to limit the validity of interview response, researchers have investigated the use of interviewers "matched" more closely to lower class respondents on a number of characteristics. As an example, the interviewing bias resulting from blacks being interviewed by whites has been well documented.[2]

Whether higher response validity from non-middle class respondents can be obtained when interviewers are matched to respondents in other dimensions such as socio-economic status, age, and area of residence has been investigated but the conclusions reached have often been contradictory.

The Los Angeles Riot Study recruited interviewers from the curfew area, i.e., the central black ghetto portion of the city.[3] However, the majority of the interviewers finally used were black housewives in the 30 to 50 age range. Although these women lived within the black community they tended to be of somewhat higher social status than most of their respondents.

Response bias was examined in detail in this study. Although there were significant differences in responses obtained by different interviewers, these differences disappeared when age was controlled. The researchers concluded that any bias introduced was not a result of personal bias of the interviewers but rather of sampling bias introduced by the interviewers due to the fact that older interviewers systematically interviewed older respondents. The magnitude of bias was small and a final conclusion was that the sampling bias exerted a "negligible effect on the total data pool."[4]

---

[1]There is some question as to the degree to which our results with the Mexican-American community can be generalized to a black situation. The problem is discussed in a study by B. P. Dohrenwend [1]. Although this study found stylistic discrepancies in responses and differing tendencies towards bias in a comparison of responses of blacks and Puerto Ricans, its conclusions suggest that our findings concerning the Mexican-American team will be of similar utility for interviewing black populations.

[2]See Hyman [2, p. 159]; Pettigrew [5, p. 50]; and Price and Searles [6, pp. 211-221].

[3]Tomlinson and Ten Houten [7].

[4]Ibid., p. 12.

A thorough study of bias when interviewing non-middle class populations has been conducted by Carol Weiss.[5] In a survey of black welfare recipients the number of truthful answers for several factual questions was obtained by comparing the responses given to interviewers with information available in existing public records (e.g., registration and voting lists, school records). Respondents with a middle class orientation were found to be more likely to bias their answers in a "socially acceptable" direction. However, the extent of the bias was influenced by the interviewer's social status and by the degree of "rapport" present in the interview (as reported by the interviewer). Both variables affected responses in the opposite direction from that which might be imagined. Interviewers of similar social status tended to receive slightly more biased responses, and interviewers reporting high rapport got much more biased responses on some questions. Rapport seemed a more important factor, which tended to either wash out or confound the bias from social status similarities. An example of the effects of the two variables is found in Table 1.

Table 1

Net Bias in Reporting Voter Registration[6]

Rapport	Interviewers Similar SES	Interviewers Dissimilar SES
High	28%	24%
Low	16	11

When obtaining information on voter registration, low rapport decreases the response bias for interviewers with social status both similar and dissimilar to that of the respondent. However, for the questions concerning school performance, interviewers of like social status with high rapport received the least biased answers of all groups.[7]

Other studies have investigated the issue of bias for non middle class samples. Williams[8] found that when a middle class interviewer interviews a lower class respondent on items which the respondent might find "threatening" (i.e., items calling for expression of attitudes or admission of behavior contrary to middle class norms), the respondent is significantly more likely to give socially acceptable responses in cases where the respondent is black and the interviewer white or when the respondent is of a lower social class than the interviewer. He theorizes that bias is affected by differences in status between respondent and interviewer, by differences in perception of norms relevant to the interviewing session, and by the degree of adequacy of the role performance, especially on the part of the interviewer.

Lenski[9] mentions two undesirable effects leading to bias when lower class respondents are interviewed by interviewers of a higher class. One is the fact that the respondent is likely to exhibit a more general reticence in this situation. The second is that the respondent may see the norms of social deference as applying even to the extent that he may passively agree to contradictory items rather than express any disagreement with the interviewer.

Further research is necessary to determine the more subtle similarities and dissimilarities between interviewer and respondent which may affect validity. Such factors might include measures of the three factors mentioned by Williams as well as measures of the similarity of attitudes and values of the interviewer and the respondent which cannot be inferred from straightforward measure of social class differences.

In reviewing the studies mentioned above as well as others, the conclusion is reached that while there are definitely variables in operation which are affecting response validity, they have not been adequately conceptualized or measured to allow for the formation and testing of hypotheses sufficient to explain the bias found.

After considerations of the pitfalls inherent in the use of either middle class or indigenous interviewers for poverty populations, the decision was made to use carefully trained indigenous interviewers.

3. THE NEIGHBORHOOD SURVEY TEAM (NST)

As part of a RAND research effort to develop methods for the evaluation of manpower training programs, the decision was made to gather new data in an experimental manner from several local manpower training programs. RAND's use of a Neighborhood Survey Team (NST) composed of indigenous staff was based on the desire to examine the following hypotheses:

1. That closely matching the respondents and interviewers on such variables as age, social class, and residence when interviewing a non middle-class population might largely avoid obstacles such as inaccuracy of response that middle class interviewers from outside have often encountered in the ghetto.

2. That using people indigenous to the ghetto neighborhood in the search for respondents would decrease the high nonresponse rates that ordinarily plague surveys in these areas.

---

[5]Weiss [8].

[6]Ibid., p. 64. The net bias is the measure which indicates responses biased in a "socially acceptable" direction. This measure is constructed by taking the total number of cases where the interviewer obtained a response different from that in the public record and subtracting those cases where the respondent reported "non-socially acceptable" behavior (such as not being registered to vote) when this was not true. The response biased to present the respondent in a more socially acceptable way was that hypothesized to be affected by the variables of rapport and similarity in social status of respondent and interviewer.

No measure of statistical significance are reported. Although the exact N for this table is not given it is probably around 500 cases.

[7]Ibid., p. 65.

[8]Williams [9].

[9]Lenski and Leggett [4].

3. That using interviewers closely matched to respondents would help the researchers understand better how to elicit and evaluate information obtained in the survey.

4. That as a by-product of this project the training received by the indigenous interviewers would increase their understanding both of their own community and of the goals of research, possibly stimulating them to put some of this new-found knowledge to use in later pursuits.

5. An additional instrumental goal was to test a method of group organization which would promote a high level of commitment and involvement, a low absenteeism rate, a low degree of friction among team members, and a low drop-out rate for the team as a whole, overcoming many of the disappointments that have occurred in past efforts to organize ghetto residents to perform tasks usually believed to require extensive training.

RECRUITMENT AND COMPOSITION OF THE NST

The NST was composed of age, sex, ethnic, social class, and residential peers of the respondent group of past participants in an anti-poverty job training program. This program is a fairly complete manpower program offering a variety of services -- general counselling, basic (remedial) education classes, prevocational classes, Neighborhood Youth Corps job slots, placement in skill training in other programs, and job referral -- to ghetto youths aged 16 to 22.[10] Team members were recruited by asking staff persons in community agencies to recommend young persons of average or above average intelligence who resided in the target area, belonged to the ethnic group predominating, and who were considered articulate, literate, responsible, and assertive enough to conduct interviews. These nominees were contacted and interviewed by a young black nonprofessional who proved quite skillful in establishing rapport with young adults and selecting good team members. This point is given some emphasis since it is commonly believed that professionals or other specially trained persons are necessary for the selection of interviewers; however, only one of the 20 team members chosen was terminated for disciplinary reasons. Most team members were recruited through the Community Service Organization (CSO), a local self-help Mexican-American group. A number of the youths had previously participated in an OEO-funded summer project through CSO, which included some unsophisticated neighborhood survey work. Half the team were males and half females, ranging in age from 17 to 24. The average age was 18. Most were high school students; the remainder were dropouts and part-time junior college students.

[10]Although our detailed finding concerning the effects of this program on the later job behavior of its participants are not yet available, a brief summary report has been issued. Holliday [3].

TEAM OPERATIONS

The team operated week nights from 4:30 to 8:30 or 9:00 p.m. and on Saturdays from 10:00 a.m. to 3:00 p.m. Preliminary data showed that most ex-enrollees of the project could not be found at home during normal working hours, regardless of their employment status.

The interviewers worked in pairs (generally one boy and one girl), for reasons of security on the ghetto streets at night and to facilitate the interviewing process. (If one member found communication with the respondent difficult, the other generally would not.) One person conducted the interview while the other recorded responses -- allowing for a smoother, more spontaneous interview. It was also hypothesized that interviewing in pairs would assist the interviewers in maintaining a sufficiently professional interviewing atmosphere and in this manner reduce bias associated with the establishment of excessive social rapport.[11]

Team members received two dollars an hour; they were not paid on a piecework basis. Typically, they assembled in the afternoon. Some immediately went out for prescheduled interview appointments; on days when interviews had not been scheduled for the entire team, members who had no appointments might go out to hunt for respondents who had no phone. Most team members averaged around 20 hours of work a week.

Absenteeism, tardiness, and other manifestations of irresponsibility remained at a low level throughout the project.

The NST operated geographically in the heart of the target area. We were offered and accepted free office space in an agency centrally located in the area under study. While economical, this windfall proved to have certain drawbacks. Interaction between agency staff and the NST tended to distract both and the high noise level and poor acoustics common to ghetto agencies hampered meetings and telephoning for interview appointments.

After some experimentation, we found that, for optimum effectiveness, an NST should have a team leader and a clerk. For this team, the leader was a nonprofessional skilled in interpersonal relations and able to relate well with young people from the ghetto. The leader became familiar with the purpose of the questionnaire and its contents as well as with the filing, paper-flow system, and payroll procedures. The clerk, a team member under the direct supervision of the team leader, was responsible for routine clerical work, the scheduling of interview appointments, and utilizing files, directories, and other sources of information on the current addresses and phone numbers of prospective respondents.

[11]Hyman finds that interview validity increases with increased "task involvement" (i.e., commitment to producing good interviews for the sake of research goals). Validity does not increase with high social involvement in which the interviewer establishes high social rapport with the respondent for other than research goals (the "hen party" situation.) Hyman [2, pp. 138-150].

The leader and the clerk worked in the neighborhood base of the NST and were on call to answer interviewers' questions and coordinate team activities.

The roles of team members and the informal leadership patterns of the team were not imposed from above but grew spontaneously from the team members themselves.

TASK ORIENTED SENSITIVITY SESSIONS

The cohesiveness and dependability of the NST appeared to be largely a function (apart from good team selection) of the deliberate encouragement of a group interaction process in which supervisor-employee relationships were informal and candid. In group discussions which might be described as "task oriented sensitivity sessions," personal and administrative problems were openly discussed. Racial and economic concerns, behavior of supervisors, and questions of RAND's and OEO's motives were also topics of discussion -- breaking taboos sometimes found in organizations, survey or otherwise. By handling these issues openly, the loyalty and esprit de corps of the team were enhanced. The interviewers were encouraged to offer comments and questions regarding the questionnaire and the interviewing process. This provided both a valuable source of feedback to the professionals and increased the interviewers' sense of participation and of their own stake in the overall project.[12]

The many sources of bias (for example those resulting from interviewers "labeling" participants in either a favorable or an unfavorable way) were examined, and the ways in which an interviewing pair could guard against such tendencies were explored. Although it was stressed to the interviewers that their ability to break through and establish rapport with respondents was crucial, commitment to the larger task goals of the research project was continually reinforced.

The organization of the interviewers into a team and the frequent, open discussion among team members and between the research staff and the team helped immensely to foster understanding and commitment to abstract research goals. This procedure would appear to be especially crucial when employing young ghetto residents who would be less likely to have developed task involvement through previous jobs in the area of interviewing. It is likely that less time would be needed to discuss with professional interviewers the goals of a research project. However, in any study requiring a group of interviewers to function as a team, as well as in survey operations where personnel conflicts arise, these group discussion sessions could prove useful.

TRAINING OF THE NST

Training began with a brief presentation of the nature and goals of the RAND research project. The presentation was forthright and informal, and questions were encouraged. Following this, there was a discussion of the training program to be investigated from the viewpoint of the team members[13] and in terms of the goals of evaluation research. A considerable period was devoted to "icebreaking" -- getting acquainted and stimulating informal interaction between participants. The proposed questionnaire was distributed, read through, and discussed. An item-by-item "run through" of the questionnaire, which in this study was constructed to obtain primarily factual rather than attitudinal data, was undertaken so that the interviewers could gain familiarity with it. The initial orientation-training sessions ran from three to four hours.

The next step was to organize the team into pairs. Some of these pairs remained intact throughout the project, while others were reassigned in light of experiences after actual field work began. The most effective method of teaching interviewing proved to be actual interviews between pair-partners in dry-run sessions. After such an interview, a debriefing was held in which a staff member and a pair discussed the interview item by item.

Accuracy, clarity of response recording, and completeness were checked and reinforced. Questions were encouraged in order to clarify concepts, develop techniques to overcome reluctance to respond, and to learn non-directive ways of clarifying or restating questions.

During the initial interviewing period a close check was maintained on interviewer performance, especially in filling out the questionnaire correctly. Continued group and individual training sessions were held as required.

INTERVIEWER TRANSPORTATION

It was discovered that many of the interviewers were not proficient at first in using street maps and guides to locate respondents, but they learned quickly through experience. It was the usual practice for two or three pairs to ride in one car to the general area in which a group of respondents were located. Here again, a certain amount of time was spent in training the team member-driver to plan his trip efficiently.

In target areas like East Los Angeles, where distances are great and public transportation is very poor, it is essential that roughly half the team members possess reliable cars in order to avoid unreasonable travel time to interviews. This may be a less troublesome problem in cities that have high population density and good public transportation.

---

[12]The Los Angeles Riot Study used somewhat similar techniques to discuss issues concerning the research project and the job of interviewing. "Human relations groups" were held on both a formal and an informal basis during the study. Tomlinson and Ten Houten [7, p. 8].

[13]As a number of team members had been at least minimally exposed to the training program, it was felt to be important to discuss this issue thoroughly. The possibility of any bias resulting from this exposure was fully discussed and hopefully controlled.

THE OPERATION OF THE NST VS. THE USE OF PROFESSIONAL INTERVIEWERS

The general operation of the team was markedly different from the operation of a survey group using professional interviewers. Due to the youth of this group and the initial lack of commitment to research goals, the interviewers were consciously welded into a group to facilitate reinforcement of the new norms and attitudes to which they were being exposed. The task oriented sensitivity sessions were also aimed at achieving this goal. These techniques might prove useful for any survey operation where young and/or inexperienced interviewers are employed.

TYPOLOGY OF TEAM MEMBERS

After a period of time, the youths on the interviewing team seemed to fall into three general categories.

The first group was composed of ideologically oriented youths who were strongly motivated to make a positive contribution to their community. This group generally formed the leadership of the team. They were interested in political and social issues, held many of the same questioning attitudes found among more middle-class youths on subjects such as war, pacifism, and racism, and were anxious to discuss broad issues at abstract levels. They contributed to the questionnaire by suggesting hypotheses, interpreting the responses of the sample to the interviewing situation, and questioning the validity and utility of the entire enterprise in a manner which proved most insightful for the research staff. Over time this group seemed to grow in influence, and their mannerisms and ideas were to some extent copied by the rest of the team. They showed the greatest amount of energy and enthusiasm of all team members but did not have the highest productivity of interviews. This group seemed proud of their ethnic identity and at the same time were able to identify with militant segments of the non-Mexican-American population.

A second group was composed of more task-oriented youths. This group comprised the "workhorses" of the team, bringing in more than their share of the interviews. This group did not seem as comfortable with their minority-status identity and often exhibited typical "anglo" (i.e., non-Mexican) tastes, patterns of speech, and behavior. Members of this group contributed to the questionnaire in a very specific way by pointing out logical inaccuracies in the unrefined early editions of the questionnaire. These contributions indicated their serious approach to interviewing per se, and their contributions seemed nearly always to arise from difficulties they might have encountered in categorizing the responses of a given respondent rather than from ideological or political principles. The task-oriented group seemed to move by choice to their "workhorse" positions and accepted some banter about this display of responsibility from the other team members. These youths seemed to be slightly afraid of being identified by the researchers as typical of the people they were interviewing. They were, for example, less likely to admit to speaking Spanish.[14]

A third group contained a number of youths who seemed to be unsure of their identities. They had not aligned themselves with either the ideologically oriented group (whose identity revolved around a positive minority image and an identification with disaffiliated elements of the non-Mexican youth culture) or the more traditionally "middle class" oriented group. In general, this group seemed more apathetic and uninvolved than either of the other two. The three types were distributed equally across sex, age and educational groups.

The fact that group members fell into these categories holds implications for team recruitment, organization, scheduling of interviews and control of bias. Both the first and the second group mentioned provide important inputs to team functioning. Future research should include systematic tests of bias in responses to attitudinal and factual questions obtained by interviewers in the different groups.

3. CONCLUSIONS

EVALUATION OF THE NST AND SUGGESTIONS FOR FUTURE RESEARCH

1. Team members, by their own report, established good rapport with respondents. They claimed they were able to sense when a respondent was not completely candid. They remedied this situation by a direct confrontation of the respondent with their awareness of his evasion and a plea to begin again and give honest answers. Task involvement was stressed with the interviewers in order that rapport might contribute to increased rather than diminished validity.

In order to test possible bias resulting from the use of the NST as interviewers, reinterviews were made on 20 randomly selected cases by a different type of interviewer, the team supervisor, at this time a non-Mexican in her mid-twenties. Responses were obtained to a subset of questions from the original questionnaire. The responses were virtually identical on the factual questions, with the exception of questions dealing with police contact, perhaps the most "threatening" asked in the reinterview.

The hypothesis that more accurate information was given the NST was supported by the fact that more respondents admitted police contact (N=6) to the NST. Three of these respondents denied police contact to the team supervisor on the second interview. By contrast there was no changed response among those reporting no police contact the first time (N=12).

Replies on the attitude questions, however, were less consistent and differences in response

[14]Through discussion sessions, the information emerged that all members of the team spoke sufficient Spanish to conduct an interview in that language.

did not exhibit any discernible patterning. An examination of the differences in response to these questions indicated that rather than tailoring their answers to present a more conventional facade to the supervisor, the respondents were simply unable to accurately recall what their attitudes had been at an earlier time.

More research is desirable in the area of bias as it occurs in factual and attitudinal questions. For the factual questions the method of the Weiss study seems most appropriate.

Further investigations of the effects of status and personality variables on bias of interview results as well as differences in response to the interview situation as a function of race or ethnicity are also necessary.

2. Although we were satisfied with the quality and the lack of bias in the interviews obtained by the NST, and with the fact that the interview refusal rate was extremely low (about 1%), the team as a whole did not demonstrate great expertness in the initial location of respondents. Part of this may be attributed to the lack of proper organization by the research staff. Although the team was well suited for obtaining information once the interview began, it would have profited from the addition of a separate group of slightly older youths who could have been regularly sent to locate the hard-to-find cases. While it is important that the respondents does not identify the interviewer with the usual bureaucracies (e.g., police, welfare, collection agencies), nevertheless a person appearing to be of slightly higher status (someone to be respected but still a compatriot) may find it easier to secure initial entry into the interviewing situation. The one slightly older male (aged 24 years) on the team was excellent in tracking down hard-to-find cases by making contacts in the neighborhood.

Future studies might experiment with the advantages and disadvantages of a variety of other techniques to increase tracking ability. Interviewers interested in tracking might receive piecework pay for locating difficult cases. The use of slightly higher status trackers or even professional tracking agencies might be tried. If the respondent's address is known at one point in time, a system of postcards for reporting changes of address, perhaps offering monetary rewards, might prove successful.

3. In addition to the main research goals discussed above, the NST did offer significant training and interesting employment to most members. Informal evidence indicates that interest in community issues seemed to be stimulated in nearly every case. At the end of the interviewing period, one interviewer announced that he had constructed a questionnaire on attitudes about prejudice, expanded from our questionnaire, and was administering it in his neighborhood in order to write a paper for his high school civics class. Other team members were placed in contact with various funding programs that make available scholarships and loans for local colleges and universities, and a number of the group expressed the desire to become social scientists. Later contact with group members indicates that most are continuing their education, and that many are moving into positions of community leadership.

The names of ex-team members have appeared in the media as being active in various community organization efforts. Team members are quick to give credit to their experience on the NST as being instrumental in helping them to move in these directions.

Our experience with this team suggests that future teams might be fruitfully organized on a semi-permanent basis, making themselves available for many different interviewing jobs in ghetto areas. Existence of readily available teams would greatly decrease research start-up costs.

4. Team members obtained a vast array of additional information that respondents volunteered during the course of the interview. Some of this information resulted in development and modification of later editions of the questionnaire. Discussions with team members proved to be a valuable source of insight to the research staff.

5. At the instrumental level, the techniques of management employed resulted in maintenance of a high level of group commitment and interest during the entire course of the project. The continuing use of the informal discussion sessions between the staff and the team was crucial in this respect.

6. As this effort was managed, the operation was not low-cost. One reason for the high cost was the long start-up time required, which was to a large extent a function of the inexperience of the research staff as well as of the NST. For example, higher costs were incurred due to the fact that the hiring of the team took less time than anticipated and the perfection of the questionnaire and drawing of the sample longer. Another factor contributing to high cost was an overestimation of the number of interviewers who could be effectively utilized at one time (this fact led to a reduction in staff about half way through the project).

REFERENCES

1. Dohrenwend, B. P., "Social Status and Psyschological Disorder: An Issue of Substance and and Issue of Method," Am. Sociol. Rev., 31, 1966, pp. 14-34.

2. Hyman, H., et al., Interviewing in Social Research, University of Chicago Press, Chicago, 1954.

3. Holliday, L. P., Appraising Selected Manpower Training Programs in the Los Angeles Area, The RAND Corporation, RM-5746-OEO, May 1969.

4. Lenski, G. E., and J. C. Leggett, "Caste, Class, and Difference in the Research Interview," Am. J. Sociol., 65, 1960, pp. 463-467.

5. Pettigrew, T. F., A Profile of the Negro American, Van Nostrand, Princeton, 1964.

6. Price, D. O., and R. Searles, "Some Effects of Interviewer-Respondent Interaction on Responses in a Survey Situation," Proceedings

of the Social Statistical Section, American Statistical Association, 1961, pp. 211-221.

7. Tomlinson, T. M., and D. L. Ten Houten, "The Los Angeles Riot Study Method: Negro Reaction Survey," Institute of Government and Public Affairs, University of California, Los Angeles, 1967

8. Weiss, C., Validity of Interview Responses of Welfare Mothers, Bureau of Applied Social Research, Columbia University, New York, 1968.

9. Williams, J. A., Jr., "Interviewer-Respondent Interaction: A Study of Bias in the Information Interview," Sociometry, 27, 1964, pp. 338-352.

II

MEASURING WELFARE LOADS

Chairman, ROBERT H. MUGGE, Social Rehabilitation Service

TRENDS IN THE AID TO DEPENDENT CHILDREN CASELOAD IN ILLINOIS

Wayne D. Epperson, Illinois Department of Public Aid

In January 1965, Illinois had an ADC caseload of 55,200 families which included 60,000 adults and 201,100 children. These families were spread over the entire state however, Cook County which includes the City of Chicago was the home of 70 per cent of them. As a ratio of the population there were 58 ADC children for each 1,000 children under 18 in Illinois. This ratio varied over the state from 1 per thousand in a county in the suburbs of Chicago up to 265 per thousand in the southern tip of the state at Cairo. Cook County, itself had 86 per thousand. For the next two years this caseload declined reaching a low point of 52,200 families in December 1966. Beginning in January 1967, the number of families receiving assistance began to move slowly upward and by August of that year, was increasing at the rate of 1,000 cases per month. In January 1969, the number had reached 71,400. During this four year period then, we saw the end of a decline in the caseload that had continued since 1962, and the beginning of a period of impressive caseload growth that is still moving upward today. In fact, the July caseload stands at 76,400.

Because of the problems created by this increasing caseload, the data available to us has been studied to determine the characteristics of these ADC families as they currently exist and to compare these characteristics with previously known data. In this way, we hoped to be able to answer some of the questions raised by Legislators, Administrators and the general public.

We had two sources of information; one, concerned with the families added and removed from the payrolls each month, and the other concerning characteristics of the families receiving assistance each month. This characteristics information is updated with each change made by the caseworker and, as an indication of its currency at any given time, Illinois requires contact with the family each quarter with appropriate changes being made at that time.

This paper then, covers a period of four years–1965, 1966, 1967, and 1968. The data relating to the additions and removals from the payrolls was accumulated for each calendar year–the data relating to characteristics is available for the month of June each year.

First, let me point out that ADC families in Illinois are dependent for three basic reasons, and these primarily refer to the father of the children although they could equally apply to the mother–physical or mental incapacity, absence from the home, or unemployment. Most of our characteristics are concerned with the parent or parents who are in the home with the children and thus subject to our scrutiny.

Secondly, policy changes relating to eligibility for ADC have been made throughout this period and we have attempted to avoid those statistics that would be most subject to these changes.

Now for some characteristics data of the caseload in June of each year:

The proportion of families in Cook County, which was 70 per cent in 1965 has slowly been increasing each year, in 1968, the percentage was 72 per cent.

The non-white portion began at 73 per cent, moved up to 74 in 1966 and eased back to 73 per cent again in 1968.

The number of children per family was 3.64, increased to 3.70 in 1966 and since then has declined to 3.55.

Ten per cent of the families in 1965 had received ADC continuously for ten years or more and this percentage has steadily risen to 14 percent.

The average age of the adult payee remained constant at 34 years until 1968 when it moved down to 33 years.

The proportion of families who have always lived in Illinois was 32 per cent–this slipped downward slightly in 1966 but has since remained at 31 per cent.

The remaining almost 70 per cent of the families, living elsewhere before coming to Illinois show the following distribution which has remained almost constant for the four years:

Adjoining states of Iowa, Missouri, Indiana and Wisconsin–6 per cent, southern states–55 per cent and the balance of almost 9 per cent from all other states and countries.

The proportion of children living only with their mother has risen from 75 to 80 per cent.

The status of the children's father has changed over the four years to some extent–the leading situation is *"Father Not Married to Mother"* which remained somewhat constant at about 38 per cent,–*"Father Deserted"* has increased each year from 27 per cent to 29 per cent, *"Divorced or Separated Father"* had also increased each year from 11 to 13 per cent–*"The Father In The Home"* situation has declined steadily from 18 to 11 per cent, *"Father Deceased"* remains very constant at 3 per cent–*"Father In Prison"* has declined slightly each year from 3 per cent to 2 per cent, and *"All Other Status"* remains constant at 2 per cent.

The educational attainment of the adults in the family has increased each year–50 per cent in 1965 had 8th grade education or less–this per cent is now 43; 37 per cent had completed some high school and this has moved up also to 43 per cent–the remaining 13 per cent in 1965 who had completed high school or more, is now at 14 per cent.

The type of work experience these adults had had remained fairly constant–approximately 40 per cent had been "service" workers–"farm or other labor" accounted for 15 per cent in 1965 but the proportion has declined to 13 per cent in 1968, "Operatives" amounted to 10 per cent in 1965 and in 1968 this was 13 per cent, "All Other" work experience moved from 11 per cent to 12 per cent and the proportion with "No Work Experience" declined from 24 per cent to 22 per cent.

In terms of "Age of Children", the large group, between 6 and 16 years of age remained steady at 58 per cent, the 3 to 6 year olds dropped from 20 per cent to 18 per cent, the "Under 3 Year Olds" remained almost constant at 15 per cent and the "16 and Over" increased from 7 to 9 per cent.

From these descriptive statistics of the families on ADC for the four year period, one fact stands out clearly–the changes that have taken place have been small or insignificant. I can summarize the findings briefly in terms of the current picture–the families are somewhat smaller, a few more live in Cook County, the non-white portion is remaining constant, families are staying longer on public assistance, there has been little change in the age of the adult payee, the proportion who have always lived in Illinois has not materially changed nor has the proportion from other places, the number of families headed by a mother has increased somewhat—the status of the father indicates a large continuing proportion are not married to the mother, desertion is increasing slightly, as is divorce and separation, the proportion of fathers in the home is declining and so is the father in prison—the adults are better educated than they were—the majority have had work experience in "Service occupations" or farm or other labor and a sizeable proportion have had no work experience at all—the age of the children in the families have not changed significantly with about 2/3 of them of school age and 1/3 under school age.

Now the additions and removals:

As you may realize, a growth or decline in a caseload is the direct result of the number of families added and removed from the payrolls each month. During this four year interval, 96,200 families were added and 80,700 were removed. At the same time, when the caseload was examined in 1968, 49 per cent of the families receiving aid in January 1965 had remained on the payroll since January 1965. This means 49 per cent of the original caseload, in January 1965, were still with us and had not been included in the numbers added or removed. All of the turnover in the four years had occurred in just over 50 per cent of the original caseload.

This turnover in families on ADC does not represent new families necessarily. In fact, some of the families have received ADC on 5 or more different occasions—a parent returns home or is remarried and it doesn't work—a job is found which is later lost due to illness or layoff. My data is based on occasions not on unduplicated families.

As families are added to the payrolls, the reasons for their dependency are coded by the caseworkers. A compilation of these reasons year by year show the following information:

Total additions each year were 21,300, 19,400, 24,300 and 31,200 for a total of 96,200 for the four years. Although provisions have been made to code many reasons for dependency, 5 of these account for over 80 per cent of the total. The leading reason is "Loss or Reduction of Earnings due to Layoff or Discharge". In 1965, this accounted for 34 per cent of all additions; in 1966, the percentage dropped to 27 per cent where it remained through 1968. Next was "Loss of Support by Parent Leaving Home" with 22 per cent in 1965, rising to 24 per cent in 1966 and 1967 and dropping to 23 per cent in 1968. "Loss or Reduction of Earnings Due to Illness or Injury" amounted to 19 per cent in 1965, moved up to 23 per cent, back to 22 and was 21 per cent in 1968, "Living Below Agency Standards at Time of Application" moved from just under 4 per cent to 7 per cent during this period—which represents an actual increase of over 200 per cent in the numerical values, "Transferred From Another Program—usually local General Assistance" remained constant at about 5 per cent.

In general terms, we see very little significant changes in the reasons for dependency with the exception of "Living Below Agency Standards".

The families added to the payrolls for reasons other than "Unemployed Parent" were further examined in relation to the "Status of the Father". In this group, which constituted over 80 per cent of all families added, the single largest status of absent father was "Deserted" which amounted to 32 per cent in 1965 and remained almost constant through the period—"Unmarried Father" moved from 23 per cent up to 25 per cent, "Divorced or Separated Father" began at 20 per cent and climbed to 22 per cent, "Father in Home" situations decreased from 12 down to 10 per cent and "All Other Absent Father Status" including Imprisonment, in Military Service, Incapacitated and Other Status decreased from 13 per cent down to 11.

From this data, we see a small increase in Father Absent due to Divorce, Separation and Unmarried and a corresponding decrease in Fathers in the Home and "Other Absent" Status.

Examination of the data on families removed from the payrolls indicate first a general decline in the numbers removed in relation to the numbers receiving assistance. The last two years of increasing caseloads has 37,400 removals compared to 43,300 during the earlier two years. In terms of reasons for removal, the major occurance was "Employment of Person in Home". In 1965, this accounted for 54 per cent and has been declining each year to the 1968 figure of 44 per cent. "No eligible Child in Home" moved from 6 per cent to 8, "Whereabouts of the Family Unknown" rose from 5 to 9 per cent, "Support from other benefits or pensions" held steady at 7 per cent, "Family Requested Removal" increased from 4 to 7 per cent. "Absent Parent Returned or Parent Remarried" has gradually risen from 9 to 10 per cent, and "All Other Reasons" remained constant at 15 per cent.

In summary, the last two years revealed the number of families removed from the payrolls because of "Employment of Person in Home" dropped 6,300 while all other reasons actually increased in number.

A large segment of the ADC population is missing from these statistics—according to a recent study approximately 114,000 absent fathers were involved in our caseload of 74,000 at the time of the study.

We attempted to collect some statistics in relation to them but at this time the results have not been completely tabulated.

Additional trend data is available to us and most of it is tabulated County by County within Illinois. As with this data presented today however little analysis is being made at this time.

AFDC MOTHERS' RESPONSE TO A MAIL QUESTIONNAIRE

Betty Burnside
National Center for Social Statistics
U. S. Department of Health, Education, and Welfare

Among recipients of federally-aided public assistance, those receiving Aid to Families with Dependent Children (AFDC) are by far the most numerous. In March 1969, close to 6.5 million persons received this form of aid.[1] A majority of the recipients are children. However, because there is no father in the home of about three-fourths of these families,[2] AFDC mothers are the second most numerous of all adult recipients of public assistance, outnumbered only by persons 65 years of age or older who receive Old Age Assistance (OAA). It is estimated that in March 1969 there were approximately 1.5 million mothers receiving AFDC.

Because the AFDC mother is often the only adult caring for several children,[3] she is undoubtedly the busiest, most hard-pressed for time among all adult recipients of federally-aided public assistance. Most of the other adults receiving aid are older in age or disabled or both, and thus less active. They have fewer dependents. In a 1965 nationwide study of Old-Age Assistance, the response rate to a mail questionnaire had been a remarkable 93 percent.[4] But mothers striving to cope with the problems of child rearing and of living on a low income might be relatively poor risks as mail questionnaire respondents, or so we thought when planning a nationwide study of AFDC in 1967. Our problem was how best to maximize returns.

Many different means, with varying degrees of success, have been used to increase mail questionnaire returns. These have included handwritten notes or postscripts, follow-up letters, sending additional questionnaires with the follow-ups, stamped (vs. franked) pre-addressed envelopes, airmail postage, special delivery postage, deadlines, timing of mailings to arrive late in the week, follow-up postal cards or phone calls, free samples, and money incentives. One study which seemed particularly relevant to our problem of attempting to maximize returns from AFDC mothers had been conducted in 1965 by Mackler.[5] As part of the study, mail questionnaires were sent to 100 low-income mothers or guardians of grade school children in New York City's central Harlem area, with an offer of $2.00 for questionnaire completion. A pencil and a stamped return envelope were also sent with the mailings. Mackler concluded that the difference between his response rate of 65 percent and the 35 percent response rate obtained in a similar study of Harlem parents[6] was largely accounted for by the money incentive.

Two states were to participate in the pretest of our AFDC mail questionnaire, one in the South and one in the West. Money payments to AFDC recipients are made by the state or local welfare agency, depending on whether the public assistance programs are state operated or state administered. There had been some speculation that the high rate of response to the 1965 Old-Age Assistance mail questionnaire, which was mailed by state welfare agencies, had been due in part to feelings of appreciation, obligation, or apprehension when approached by an agency which provided them assistance.[7] We in turn speculated that such feelings might be less likely to occur if questionnaires came from a federal agency such as the U. S. Department of Health, Education, and Welfare. In our pretest, we decided to test this assumption, and also to test the incentives of a pencil sent with the questionnaire and an offer of $2.00 for questionnaire completion.

The pretest was therefore designed to include eight different subsamples of equal size, four of them with federal study sponsorship and four with state study sponsorship. Under each type of sponsorship, the subcategories were a combination of money and pencil incentives, money only, pencil only, and no incentive. The pretest sample consisted of a total of 432 open AFDC cases, i.e., families currently receiving money payments, in which there was a mother or other female caretaker of the AFDC children in the home. Very few AFDC families have no adult female in the home, but we wanted all respondents to be female and thus use sex as a control variable. Each of the two participating states furnished 216 cases drawn by systematic sampling. Sample cases were assigned consecutive numbers by the states and then randomly assigned to the eight subsample categories.

The content of the mail questionnaire was based largely on data obtained by interviews in a two-state pilot study of AFDC conducted in 1965.[8] Pilot study items used were those considered to be most successful in identifying relative deprivation and in revealing respondent attitudes about welfare. For a number of questions in the mail questionnaire, including some asking for opinions, precoded answer categories were adapted from pilot study open-end question reply categories. Although aware of the drawbacks of fixed alternative opinion questions, we believed that in regard to welfare, most public assistance recipients would be likely to hold fairly clear opinions.[9] We kept the mail questionnaire short, with only 32 questions for everyone, plus nine contingency questions. We provided an "anything else you have to say about welfare" question at the end as a safety valve for respondents who might feel frustrated with the alternatives of a rigidly structured questionnaire (there were only two open-end questions) or who wanted to discuss their problems.[10] To make the questionnaire appear as brief as possible, we used a double-column format, which resulted in a single-fold, pamphlet-type instrument of four pages.

Altogether there were three mailings, with the two follow-ups spaced at 12-day intervals. An explanatory letter of transmittal, a copy of the questionnaire, and a stamped, self-addressed envelope were sent in the initial mailing and first follow-up. The second follow-up consisted only of a short letter of reminder. In the initial mailing and first follow-up letters, respondents were told that the sponsoring agency was trying to plan better welfare programs, followed by the statement, "... and we think that people who have been on welfare can help us." The women were also told that their names had been selected by chance, their answers were needed but there was no obligation to reply, and they were assured of confidentiality. For respondents in the money incentive subsamples, a statement was included that the $2.00 payment would not affect their welfare money grant. They were also asked to furnish, on a separate form, the name and address where the $2.00 was to be sent, just in case some might wish to have the money mailed to another address or to another person.

We set a deadline of sorts,[11] requesting that the questionnaire be filled out five days after its receipt. Judging from the response, this deadline was at least moderately successful. By the time of the second mailing the response rate was 63 percent. By the time of the third mailing the response rate was 78 percent. For the pretest, the total response rate was 86 percent. Only 11 of the 432 women in the sample had moved and could not be located by the Post Office.

We were elated by the response rate even though it fell short of the 93 percent rate in our 1965 Old-Age Assistance study. Rates reported for other mail questionnaire studies have usually been in the range of 10 to 70 percent.[12] But our principal interest was in rates for the different combinations of study sponsorship and incentives. Response rates for the eight subsamples were:

Subsample	Percent responding	
	Study sponsorship	
	Federal	State
Money and pencil incentives	86.0	88.5
Money only	90.2	88.7
Pencil only	90.6	84.9
No incentive	92.6	85.2

The pretest sample design included an expected response rate of 80 percent and required a difference between subsample response rates of at least 10 percent for statistical significance at the .10 level. Based on our returns, we therefore concluded that neither the money offer nor sending a pencil nor type of study sponsorship, in any of the combinations, had significantly affected response rates.

Although there were no significant differences between subsample response rates, we decided to test the relationship between the returns and several relevant study variables. One of particular relevance was current welfare status, because women no longer getting aid would presumably be less likely than those still on the welfare rolls to feel obligated to respond or be apprehensive if they did not. Cases currently open were selected for the pretest, but a time lapse of a month or so was unavoidable between drawing of the sample and the first mailing of the questionnaire, and case turnover for AFDC tends to be highest among all types of public assistance.[13] We had asked the women if they were currently receiving welfare, and 30 of the 371 who responded replied "no". Unfortunately, this distribution of the variable made it inadvisable to use a chi square test of significance, even with correction for continuity.

A comment here in regard to contingency questions is of methodological interest. Following the question, "Are you getting welfare now?" those who replied negatively were asked when they got their last welfare check. The contingency question was answered by all 30 of the women no longer getting welfare, but it was also answered by 84 percent of respondents who should not have answered it. Mention of receipt of the welfare check, an understandably important event in the lives of welfare recipients, is apparently a powerful stimulus. We believe that the response to this question makes it one of the most unsuccessful contingencies in the history of survey research--a dubious distinction. In the final version of the questionnaire, we took the easy way out and asked everyone the former contingency question; unwanted responses could be disposed of by cross-tabulation.

Study variables, other than current receipt of welfare, which seemed most relevant to the subsample returns were respondent's race (white, Negro, other races); school grade completion (0-4 grades, 5-8, 9 or more); urban or rural place of residence; state of residence; her report of either having been or never having been denied welfare at some time; and her total time on welfare, including past and present episodes (under 1 year, 1-3 years, more than 3 years). With the listings, the two states had furnished us data concerning race and urbanization of place of residence; other data were obtained by the questionnaire. Using the .10 level with the chi square test, as in the study design, the only significant relationship found was that between subsample returns and respondent's total time on welfare. Even though we could not predict the direction of the relationship, the finding appears to warrant further research at some time in the future. It is noteworthy that the chi square test result for relationship between returns and state of residence was unusually small.

CONCLUSIONS

Based on pretest results, we concluded that in our nationwide AFDC study we could expect a fairly high response rate without the use of incentives. We did choose federal rather than state study sponsorship for several reasons, including the slightly higher total response rate for federal mailings (87 percent, compared with a state rate of 85 percent), ease of conducting the study from one central location, and the fact that there was some evidence, although not conclusive, that certain types of open-end replies would be made to a federal but not to a state agency. The relatively few criticisms of welfare workers and allegations of discrimination by respondents had all been made in questionnaires mailed to the Department of Health, Education, and Welfare. We decided to repeat the use of two follow-ups and stamped, self-addressed return envelopes because we believed they had probably had a favorable effect upon the response rate although they were not specifically included in the pretest design. Findings from other studies which have tested these variables indicate that they are effective as inducements to response.[14]

In our nationwide study, the mail questionnaire was sent to a representative sample of 3,659 mothers or female caretakers of AFDC children in the conterminous United States. A total of 2,969 women returned completed questionnaires, for a response rate of 81 percent.[15] Although this was somewhat lower than the pretest response rate of 86 percent, it was relatively high for a mail questionnaire survey. Our pretest results were therefore substantiated: Most AFDC mothers, although preoccupied with the problems of child rearing and of living on a low income, would take the time to reply to a mail questionnaire without having received any incentives. There were expressions of appreciation in open-end replies, there may have been feelings of obligation or apprehension on the part of some respondents, but we believe that many of the women were also genuinely interested in trying to provide information which they felt would help the government to plan better welfare programs. In other words, as formulated by Kahn and Cannell,[16] recipients of AFDC were motivated to respond because of their perception that by communicating they would move toward certain of their own goals. We know with certainty that at least some of our respondents felt this way for in addition to answering the structured questions, they wrote in specific recommendations for changes in the AFDC program.

FOOTNOTES

1. Public Assistance Statistics: March 1969, U. S. Department of Health, Education, and Welfare, Social and Rehabilitation Service, National Center for Social Statistics (NCSS Report A-2 (3/69)), table 7.
2. Preliminary Report of Findings: 1967 AFDC Study, U. S. Department of Health, Education, and Welfare, Social and Rehabilitation Service, National Center for Social Statistics, October 1968 (NCSS Report AFDC-1 (67)), table 7, p. 9.
3. Ibid., table 1, p. 6.
4. Findings of the 1965 Survey of OAA Recipients: Data by State and Census Division, Part II Data Obtained by Mail Questionnaire, U. S. Department of Health, Education, and Welfare, Social and Rehabilitation Service, September 1967.
5. Oliver C. Moles, Lola M. Irelan, and Bernard Mackler, "Use of Mail Questionnaires to Collect Data from Low-Income Families," Welfare in Review, Vol. 5, No. 2, February 1967, pp. 21-24.
6. Ibid.
7. Ibid.
8. Report on Pilot Study of Family Living Conditions, Aid to Families with Dependent Children: Continued and Closed Cases, Denied and Accepted Applications, U. S. Department of Health, Education, and Welfare, Welfare Administration, Bureau of Family Services, Division of Research, July 1967. Unpublished.
9. For a discussion of fixed alternative questions, see Claire Selltiz, Marie Jahoda, Morton Deutch, and Stuart W. Cook, Research Methods in Social Relations (New York: Henry Holt and Company, 1960), pp. 258-262.
10. A device suggested by Sol Levine and Gerald Gordon in "Maximizing Returns on Mail Questionnaires," Public Opinion Quarterly, Vol. 22, No. 4, Winter 1958-59, pp. 568-575.
11. Another device suggested by Levine and Gordon, op. cit.
12. See, for example, Selltiz, et al., op. cit., p. 241; William J. Goode and Paul J. Hatt, Methods in Social Research (New York: McGraw-Hill, 1952), p. 173.
13. Net change in caseload per 1,000 cases open during year: AFDC (53 states) +94, Old-Age Assistance (35 states) +3, Aid to the Blind (35 states) -11, Aid to the Permanently and Totally Disabled (34 states) +82, Aid to the Aged, Blind, and Disabled (18 states) +12, General Assistance (30 states) +56. Public Assistance Programs: Indices of Case Turnover, Calendar Year 1967, U. S. Department of Health, Education, and Welfare, Social and Rehabilitation Service, National Center for Social Statistics. Unpublished.
14. See, for example, William M. Kephart and Marvin Bressler, "Increasing the Responses to Mail Questionnaires: A Research Study," Public Opinion Quarterly, Vol. 22, No. 2, Summer 1958, pp. 123-132, Stanley D. Bachrach and Harry M. Scoble, "Mail Questionnaire Efficiency: Controlled Reduction of Nonresponse," Public Opinion Quarterly, Vol. 31, No. 2, Summer 1967, pp. 265-271.
15. 1967 AFDC Study: Preliminary Report of Findings from Mail Questionnaire, U. S. Department of Health, Education, and Welfare, Social and Rehabilitation Service, National Center for Social Statistics, January 1969 (NCSS Report AFDC-2 (67)).
16. Robert L. Kahn and Charles F. Cannell, The Dynamics of Interviewing (New York: John Wiley & Sons, Inc., 1957), p. 121.

DISCUSSION

Elizabeth Durbin, New York University

Before making specific comments on the papers we have just heard, I should admit that I have a certain prejudice against surveying welfare clients yet again, particularly when the explicit or implicit purpose of the study is to say something about why welfare caseloads have been increasing. The choice of surveys as the method for answering questions about welfare arises, I think, from the client orientation of many of the welfare researchers, themselves from the social work tradition. In other words it reflects what I consider a mistaken point of view, that the answers as to why there are increasing numbers of people on welfare must lie in some attribute of the recipients themselves. I have even had the impression that there is an unspoken fear that if it's not something to do with the clients, then the only alternative explanation must lie in some failure on the part of social workers. Too often welfare administrators and researchers retreat into the position that we don't know enough about clients to explain caseload increases, rather than to ask, <u>first</u>, what is it we need to know to understand caseload increases, and then what are the appropriate means to go about getting answers.

Indeed besides my prejudice against surveys, I am also not really competent to assess the methodology of any particular survey. So Miss Burnside will forgive me if I confine myself to a very few remarks on her paper. I was impressed with the high response rate she reported. Since the purpose of the pilot study was to find the best means to maximize national returns, it would seem that the object was met. However it would have been interesting to know how useful the pretests were in explaining the somewhat lower national response rate. For instance, was the response rate lower in all parts of the country, or did particular areas bring the average down? And if so, were there any explanations for such differences? In reading Miss Burnsides paper, I was also sorry from the methodological research point of view that she hadn't broadened the concept of the incentive to respond. Particularly if she felt that an important reason for the high response rate was the motivation to contribute to changes in welfare policies, it would have been interesting to have tested differences in the wording of that introductory letter. But I realize this is being somewhat wise after the event.

Since I have been involved in a study of welfare caseloads in New York City, and have collected data similar to Mr. Epperson's, I have more experience to draw upon in commenting on his paper. As my introductory remarks also implied I do feel that there are more fruitful approaches to understanding the dynamics of welfare caseload than surveys of recipients. Indeed it is precisely the understanding of the inflows and outflows on welfare programs, to which Mr. Epperson's paper contributes, that is vital for us to begin to understand why families have been seeking government income support in ever increasing numbers. For this reason I was also particularly disappointed that Mr. Epperson refrained from making any interpretations of the data he presented.

My findings in New York City were similar to much that Mr. Epperson found in Illinois. Although I have separate data for ADC and the unemployed part of the program, I also found that loss of a job, loss of a parent and illness were the primary reasons given for opening cases. Employment was also an important cause for closing ADC cases, although the reason "increase of resources" was as important in New York City. However it would seem that in the closings data the reasons are classified somewhat differently in New York from Illinois. Comparable figures on the status of mothers and fathers show that a larger proportion of mothers coming on ADC are unmarried in New York, 32%, and desertions considerably higher than divorces and separations, possibly due to some difference in the process of getting separations. I also found that only about 12% of ADC homes had fathers present. You may be interested to know that I collected such data for ten years, and over the longer period considerably greater changes had taken place than in the period 1965 to 1969, both with respect to reasons for opening and closing cases and for family status.

However, I dispute Mr. Epperson's conclusions that changes have been "small or insignificant", on conceptual grounds rather than on the basis of his data. Indeed I would contend that his data suggest the opposite of his conclusions. For while his data on the characteristics of the total caseload show unmarried mothers at 38%, desertion at 27-29% and divorce or separation at 11-13%, the data on incoming cases show a much lower proportion of unmarried mothers, a somewhat higher proportion of divorced and separated mothers. This suggests to me at any rate that the cases now opening are considerably different from the average, and thus in time would be expected to alter the average characteristics in this direction. This demonstrates the general proposition that small changes in characteristics may conceal important changes in inflow; for the inflow is generally fairly small compared to total caseload. And incidentally this is a good illustration of the deficiencies of surveys of the whole caseload in understanding current dynamics.

Finally while I applaud the generation of the kind of data Mr. Epperson presented, I do feel it is somewhat inadequate for his stated purpose of answering some of the questions raised about rapidly increasing caseloads. For the truth of the matter is that caseloads have been rising for a number of reasons, not least of which is that welfare administrators, for one reason or another, have been letting many more applicants on to the programs than hitherto. I do not know the situation in Illinois, but in New York City applications for ADC increased from roughly 4,000 a month in 1965 to 6,000 in 1967, but openings almost doubled, reflected in the increase in the acceptance rate from 64% to 76%. Another important factor in New York State was the 30% increase in benefit allowances in 1966, which automatically increased the eligible population. The basic problem of any adequate analysis of caseload increases is to isolate on one hand changes due to increased availability of welfare, which may be manifested in increased acceptance rates, rising allowance levels or both, from, on the other hand, the responses of the potential client population to these and other relevant factors. We are only just beginning to think about how this can best be done. Thank you.

Discussion
Stephen Leeds, National Self-Help Corporation

It seems to me that there are three major factors which must be dealt with in "measuring welfare loads," if this measurement is to have any meaning. Since welfare is a very important contemporary issue, it does not suffice to rattle off some static indicators of caseload characteristics, nor to test some laboratory-like hypotheses concerning response rates to mailed questionnaires. Furthermore, asserting that the AFDC rolls have evinced a meteoric rise in the last three years despite great economic progress reveals a misunderstanding of what welfare really is all about. Welfare is a measure of imbalance in our society.

The first factor with which we must deal is related to the dimensions of the poverty pool of eligible and potential recipients, and thus to the economics of race and class in America. Data concerning this issue are available, or may be derived, from various Bureau of Labor studies from the Census, and from many surveys. We must confront the fact that the median family income in New York City's ghettos was $4000 in 1966, or exactly what it was in 1960....except infortunately for inflation and sales taxes, etc. And we must confront the fact that one out of every five jobs in this City today pays under $80 per week, that minimum wages stand at $3200 per year -- gross income -- while the current welfare grant for a family of four is roughly $3500 annually -- net income!

Yes we've had economic progress, but not the poor. They are worse off today than six or eight years ago. Approximately 200,000 families in New York City are headed by a man who is working but earning less than what welfare defines as his household's minimum needs. Most of these families are Black or Puerto Rican. Why should a man work 40 or more hours each week to maintain his family in defined poverty? Out of pride? Why shouldn't the man leave, at least ostensibly, keeping his $3200 job, and put his wife and children on welfare for another $3000, or more....thus bringing his total family income close to the median for the City as a whole of $7000 to $8000?

The second factor in measuring welfare loads is related to these questions: acceptability of welfare. Welfare, for many, is no longer a dirty word. In fact it's not a problem, but rather a solution. Welfare has become a "right;" it is no longer a privilege. Get out your attitudinal surveys and delve into the Black or Puerto Rican mind. You know, in this City about 30% of the Blacks and 40% of the Latins receive welfare. In the face of such staggering statistics it must be obvious that welfare is now acceptable. Look at the newspapers; recipients are organizing and demonstrating for benefits. They are not ashamed or embarrassed! Welfare is their right. But more than that too. Anti-poverty agencies, community corporations, and Medicaid teams have played no little part in referring potential clients to the Department of Social Services. Welfare is acceptable to many of the professional and "para-professional" people in the poor communities. It is no coincidence that the caseloads doubled with the advent of the War on Poverty, and with the incursion of these agents into the ghetto's daily life.

The final factor I wish to bring to your attention is perhaps the most difficult to measure: the political-administrative welfare apparatus. Let me give an example. The State Legislature passed a regressive welfare bill here on March 29. While grants were cut by about 8%, and although such humane concepts as minimum standards were abolished, this Bill should not have had too great an effect on caseload growth. Poverty was still here; inflation was still rampant; welfare was still a right. Yet the word was out; the screws were being tightened. Almost immediately, the caseload growth dropped from an annual average of 16,000 additional persons each month to only 5,000.

In what can only be characterized as a "work action," the Department of Social Services began to follow the letter of the law with regard to welfare applications. Every document suddenly had to be verified, although the Department has maintained that its ineligibility rate among recipients was constant for years at well under 5%. Clients were harrassed, applicants discouraged. How else does one explain a 60% drop in growth rate within the course of a month or two. But how does one measure the variable enforcement of laws and regulations, and how does one measure the side-effects of crackdown and repression?

I want to be brief, but I can not close without mentioning President Nixon's recent welfare proposal, especially that part related to Manpower and Training, which is close to my area of expertise, and which I consider doomed to abysmal failure.

New York City's Department of Social Services schedules some 10,000 interviews each month with recipients to determine employability. Of these, 5000 are kept. Of these, some 500 result in job or training placements; for a 5% to 10% batting average. Four out of five jobs obtained for recipients pay less that $80 per week, in the most vulnerable occupational areas. And former welfare recipients act no differently than other people in dead-end jobs: they lose them. A child-care problem, a sudden illness, or a moderate debt -- these are enough to create a crisis which puts the family back on the dole. With no financial cushion, they have nowhere else to turn.

So we find, in New York City, that almost 40% of the AFDC case openings each year are in fact re-openings. Of the roughly 20,000 AFDC case re-openings each fiscal year which closed previously during the same fiscal year, about 12,000 are brought about by job loss. There is a revolving door between welfare and work among the poor. In fact, some 15% or more of this City's recipients are in families getting public assistance in addition to wages. Many other cases are publicly aided in addition to work-related benefits such as OASDI, pensions, unemployment compensation, etc.

A study I conducted of male heads of AFDC-U households in New York revealed that less than one third were rated as having good potential for more than marginal jobs. Almost half were in need of intensive remediation; over two fifths had less than a grammar school education; one fifth had chronic health problems; and half manifested a history of poor job stability.

Let's face facts! Many welfare recipients are not going to become permanently self-supporting at a decent standard of living under our current system. In its first year of operation, New York City's Employment Incentive Program, which allowed clients to keep a fair proportion of earnings in addition to their welfare grant, only managed to remove 100 of the roughly 4000 participants from the rolls. The Work Incentive Program in this City has only filled about 3000 of its 8500 slots; and this includes both training and employment opportunities.

So what's all this talk of Manpower and Training. People work for money in this country -- for how much they can go into the weekend with. People are not going to go out of their way to earn $3920 a year, at least not up North. We don't have the jobs and we're not willing to pay and we're not going to put a dent in the rolls under Nixon's plan.

Is there a way out? Yes! You know, in New York City garbagemen earn $9500 per year, base salary. Here's a job that requires no training, no skill, but pays almost $5 an hour. The government foots the bill. If the government were willing to use the same criteria for other jobs as it does for garbagemen, it could lick the welfare problem. The secret is adequate pay, a guaranteed adequate income. If we offered every AFDC mother, and every working man with a family, a job paying $7000 or more we'd cut the rolls by 70%.

III

HOUSEHOLD SURVEYS OF EMPLOYMENT PROBLEMS IN URBAN POVERTY NEIGHBORHOODS

Chairman, HAROLD GOLDSTEIN, U. S. Bureau of Labor Statistics

METHODOLOGICAL AND INTERVIEWING PROBLEMS IN HOUSEHOLD SURVEYS OF EMPLOYMENT PROBLEMS IN URBAN POVERTY NEIGHBORHOODS

Earle J. Gerson, U. S. Bureau of the Census

Description of the Survey

Despite the relative abundance of National data relating to poverty and employment problems in the United States, the amount of current information available for specific urban poverty neighborhoods has been negligible. Although many persons had called for study and several major action programs had begun earlier, the Nation's attention began to focus on problems of the urban poor after the Watts riot of 1965. At that time, the Census Bureau was asked to conduct a Population and Housing Census of the area to provide officials with an up-to-date inventory of social and economic conditions.

Since then, several Federal agencies, including the Department of Labor, the Department of Housing and Urban Development, and the Office of Economic Opportunity, have established programs to combat the perceived problems of urban poverty neighborhoods. During the development and operation of these programs, it became clear that information more current and in some respects more detailed than decennial Census data would be valuable for efficient allocation of program resources. Several National poverty-oriented studies were undertaken during this period, including the Survey of Economic Opportunity, sponsored by the Office of Economic Opportunity, and longitudinal surveys sponsored by the Department of Labor and the Social Security Administration. However, these did not meet the needs for data for individual poverty areas.

Accordingly, in July 1968, the Department of Labor initiated the Urban Employment Survey in selected poverty neighborhoods of six large cities. The Census Bureau is responsible for collecting and compiling the survey data and the Bureau of Labor Statistics, for analyzing and publishing the survey results.

Although not a case history, this paper describes the principal features of the survey, problems encountered to date, and some concerns regarding potential problems.

The survey covers a broad range of topics, some of which are covered Nationally and on a regular basis by the Current Population Survey. Other topics are specifically focused on poverty neighborhood problems and would have limited relevance to other populations.

The principal subjects covered are current labor force status, work experience, and family income, which are regularly included in the Current Population Survey, and work history, job training, migration history, ethnicity, barriers to employment, and job attitudes, which are not. Because of the importance of the topic and the need to explore each category of involuntary nonparticipants in the labor force in detail the inquiry on barriers to employment constitutes the largest single set of questions.

The required information on household membership and characteristics, and on the current labor force status of persons not present at the initial interview, is provided by a responsible adult household member. Each adult (16 years old and over) is then interviewed personally to verify the current labor force status reported for him and to provide information on the other topics covered in the survey. Approximately 15 minutes are required to complete the household questions, and 20 minutes per adult for the personal interview.

The six cities included in the UES are Atlanta, Chicago, Detroit, Houston, Los Angeles, and New York. They were selected because the Labor Department operates a concentrated Employment Program in each, and because of the ethnic and regional diversity they encompass. In each, a set of Census tracts, not necessarily contiguous, was designated as the sample area. In addition, the remainder of the city in Atlanta and in Detroit were designated as separate sample areas, in large measure to serve as control groups.

The reliability objective for each area is to produce an annual average estimate of a characteristic held by two percent of the population, with a coefficient of variation of ten percent. Accordingly, a probability sample of 3500 interviewed households (7700 persons 16 years and over) per area, per year was designated for the survey. This is divided into 52 weekly subsamples of approximately 70 households. Interviewing is spread uniformly through the year for efficiency of data collection and to control for the effects of seasonal variation. After the initial year of operation, July 1968-June 1969, one-half of the sample each week is being interviewed for the first time and one-half are retained from the previous year's sample. This partial sample rotation pattern, when compared with use of a new sample each year, is expected to yield improved reliability for estimates of year-to-year change.

The data are collected by a staff of interviewers who, for the most part, work full-time on the survey and who reside in the neighborhoods covered by the survey. They are under the direction of the Census Bureau's full-time field supervisory staff which, in five of the cities (all except Houston), is located in the same city. In three of the cities, separate offices are maintained in the poverty neighborhoods for ease of communication with the interviewers.

Each beginning interviewer receives several days of classroom training, followed by one or two days of on-the-job training. Thereafter, he periodically is given self-study assignments and further personal training and is observed by a supervisor while working. In addition, all of his questionnaires are edited and a sample of his work is reinterviewed by supervisory personnel. Corrective action is taken as required.

Because a major purpose of the survey is to meet data needs for the individual poverty neighborhoods, separate tabulations are prepared for each of the eight sample areas. Tables based on the approximately 175 items on the questionnaires are prepared for the total population of each sample area and for major ethnic groups within it. The data for individual cities are analyzed and published by the Bureau of Labor Statistics Regional Offices.

The operations and design features described above relate to the cross-sectional aspect of the survey, which describes the characteristics of the current residents of these neighborhoods. It is possible that, if anti-poverty efforts in these areas are successful, the characteristics of the current neighborhood residents might not fully reflect that success. Those who acquire better jobs and higher income through the efforts of the programs may seek better housing and a more favorable environment in different neighborhoods. They may be replaced by persons of lower economic status or the population of the area may decline.

To gain some insight into this problem, a sample of persons who were interviewed for the survey are interviewed one year later. Most will be found at the addresses at which they were originally interviewed, but others will be followed to their new addresses anywhere in the conterminous United States. The sample of persons who moved is not considered large enough to provide sufficiently reliable data for the individual poverty neighborhoods, and will be tabulated at the end of the second year of the survey for all six poverty neighborhoods combined.

Methodological and Interviewing Problems

The problems of conducting the Urban Employment Survey are basically similar to those for other household surveys, but their severity and implications may be greater in UES. To the extent that they vary in degree between cities and over time, they may also make inter-city and time series comparisons difficult.

The first problem area is coverage of the population. How complete is the sample and does it include proportionally all elements of the sample area population?

The designation of addresses for inclusion in the sample was from reasonably complete sampling frames, was carefully checked, and appears not to account for any substantial undercoverage. Any problem of population coverage there may be would, therefore, relate largely to missed persons at the designated sample addresses.

Unfortunately, there is little direct information available regarding the size and characteristics of the current population of poverty neighborhoods, as defined for this survey. Such information would be required to determine whether a coverage problem exists, how great it is, and which sub-groups are most affected.

Independent population estimates are available Nationally and are used in evaluating coverage in the Current Population Survey and in decennial Censuses of Population. The general conclusion of coverage evaluations for these programs is that population groups of major concern in a study of poverty neighborhood employment problems--Negroes and young males--present a substantial coverage problem.

Siegel (1) has prepared estimates of population undercoverage in the Current Population Survey (CPS) in 1965. Since the coverage procedures used in the CPS and the UES are similar, Siegel's estimates can be considered indicative of the magnitude of the problem that might be expected in the UES. Unfortunately, however, they relate to National population estimates and not to the central cities of large metropolitan areas that are the UES areas.

Siegel's data (2) (Table 1) were derived by comparing population estimates from the CPS before adjustment to independent population controls with a "corrected" or best estimate of the population based on demographic analysis. They indicate that the CPS, with its reasonably well trained interviewing staff and its stress on quality control, has a significant undercoverage problem and that the problem is greatest among Negroes and other races.

TABLE 1. Estimated Percentage Net Understatement of the Population 14 Years and Over, by Age, Sex, and Color, in the Unadjusted Current Population Survey Estimates: 1965

Age	Negro and Other Races		White	
	Male	Female	Male	Female
Total, 14 and over	16.8	8.8	7.0	4.2
14-19	7.3	5.2	2.8	3.1
20-24	30.1	16.5	10.9	7.8
25-29	20.9	11.7	9.2	6.0
30-34	23.7	8.3	9.3	2.8
35-39	22.6	2.3	6.8	2.7
40-44	18.1	0.2	4.3	*
45-49	15.8	6.0	6.7	1.0
50-54	13.0	9.6	6.0	4.3
55-59	18.5	20.5	8.0	6.9
60-64	13.3	7.2	7.5	6.4
65 and over	6.4	13.4	8.7	5.7

* Percentage between + 0.05 and - 0.05

Only thirty percent of the poverty neighborhood population covered by UES is white, and of this thirty percent, most are Puerto Ricans or Mexican Americans who possibly present coverage problems of the same nature as Negroes and other races. Thus, the data in Table 1 for Negroes and other races imply much regarding the UES results. If UES does miss close to one-fifth of the Negro males in the prime working years, what is the impact on such measures as family income, unemployment rates, and participation in job training programs?

At present, there is no fully adequate direct measure of UES coverage. Although a reinterview program is conducted using supervisory personnel, this method has not, in other current surveys, measured the full extent of undercoverage. Comparisons of UES estimates with data from the 1970 Census will be made, but this will provide only a measure of relative coverage.

A second major problem is that of nonresponse. There are two types of nonresponse in the UES. The first is a failure to contact and interview any adult member of a household. This household respondent, as noted earlier, provides information on the characteristics of the household, current labor force status of persons not at home during the initial interview, and all required information relating to himself. In this type of noninterview virtually no information relating to the household is available. The second type of nonresponse is failure to interview each person 16 years of age and older to confirm his current labor force status, and obtain the required data on work history, income, barriers to employment, job attitudes, and other subjects reserved for the individual's interview. Thus, in this type, most of the required information is obtained from the household members who are interviewed. Only the part of the report to be obtained from missed individuals is lacking.

Interviewers are allowed nine days to complete their assignments. Thereafter, the field supervisor personally attempts to reduce the number of noninterviews. In some cities noninterviews are reduced by about half through this procedure.

Although imputations are made for missed household and personal interviews, they are undesirable substitutes for reports from the designated households or individuals. Since several studies have indicated that the characteristics of nonrespondents often differ markedly from those of respondents, it is reasonable to conclude that the survey results are subject to noninterview bias of unknown magnitude which is proportional to the noninterview rate.

In the three months ending in June 1969, the noninterview rate for households was 4.1 percent and for persons, 5.4 percent. Although the overall household noninterview rate compares favorably with that for poverty areas in other surveys and has declined in recent months, the combined effect of the two types of noninterviews on the survey results can be important. Close to one-tenth of the data derived from the personal interview portion of the survey are imputed, based on the reports of interviewed persons. This rate varies widely by city, with the result that the imputation rate for two of the six cities is approximately 50 percent higher than the six-city average.

The major burden of a successful data collection operation falls on the interviewing staff. Reaching difficult to contact persons, evoking cooperation from the reluctant, and obtaining full and accurate responses to the questionnaire are skills that must be developed. In a long interview with the culturally deprived, perhaps suspicious or cynical poverty neighborhood residents, a higher than ordinary skill would seem appropriate.

The interviewers employed for the survey are high school graduates and some have attended college. The majority work full-time as Census Bureau interviewers although some work on other surveys, as well as the UES. Their rate of pay is approximately 20 percent higher than other Census Bureau interviewers. About 40 percent are males.

A study of learning curves for Current Population Survey interviewers has shown that two and one-half years (3) are required for an interviewer to achieve peak performance on the survey as measured by noninterview rates and the frequency of edit problems. A high rate of staff turnover, therefore, can be interpreted as an indication that there is insufficient time for the average interviewer to develop and accumulate higher level interviewing skills.

Interviewer turnover, defined as the number trained in excess of the number of positions, was over 100 percent during the first year of the survey. Of the 72 interviewer positions in July 1969, only 27 are filled by persons trained a year earlier at the start of the survey.

The reasons for the high turnover rate cannot be fully determined, but several categories of separations are indicative of the problems of optimum selection and retention of staff.

First, there are competing opportunities in the job market. Some interviewers have left for more remunerative or desirable employment. Second, others have lacked basic language, quantitative, or other skills required for adequate job performance. Next, a number of interviewers were unable to function adequately in a situation wherein they establish their own work schedule and are required to stay current with their work and meet Census Bureau performance standards.

Another reason for separation is a reluctance to be exposed continuously to the crime problem in poverty neighborhoods. Several interviewers have been beaten and robbed while interviewing.

Finally, a few were dismissed because the reinterview program disclosed that they had falsified some interviews. Falsification is not an uncommon problem in surveys. Steinkamp(4) noted that three of twenty-one interviewers used in a Consumer Savings Project were found to have "curbstoned." Gallup (5) was interviewed by the New York Times after curbstoning by two of twenty-six interviewers had been discovered in an attitude survey of Harlem residents. He stated that "the difficulties of doing a scientific poll in Harlem are extreme ... [and that] a few other ghetto districts might be equally tough .... The normal living patterns are completely disarranged .... They just don't want to talk to a stranger."

The importance of the interviewer's role in the communication process and the quality and money costs of turnover, make this one of the major problems of the program.

The final major category of methodological problems relates to the comparability of UES data with the results of other surveys and Censuses. Certainly, comparisons of data from the UES and the Current Population Survey will be made.

Even with the use of identical questionnaires in two surveys, differentials in the degree of error or in the procedures followed can account for some of the differences in the results. In addition to the problems of coverage and noninterview bias cited earlier, the relative frequency of response error and processing error can have a major effect on comparability. It is seldom possible to measure the total error in surveys adequately, but there is reason to believe that for some subjects it is of sufficient magnitude that sharp variations can effect the comparability of survey results.

There are two notable aspects in which UES procedures differ from those in CPS which may reduce comparability. While not specifically problems of conducting surveys in poverty neighborhoods, they do reflect on possible uses of the UES results. They are also indicative of the concerns of the survey practitioner in designing a survey program, the results of which will be linked with other programs.

One is the sample rotation pattern. In CPS, households are interviewed four times at monthly intervals, are dropped for eight months, and then returned for four more monthly interviews. There is a pattern of differences in unemployment rates by month in sample. the UES, on the other hand, retains sample households for only two interviews, scheduled one year apart. If there is a differential conditioning effect in the two surveys, it will produce spurious differences in the data. This effect in the UES data will be studied as the second year results become available.

The other example of noncomparable procedures is the use of a household respondent in CPS and a direct interview with each person in UES. Although the effect of this might vary for the many topics covered in the surveys, it cannot be disregarded.

The separate effects of several of the categories of problems mentioned here can, in time, be measured or at least estimated. They have been presented here primarily as illustrations of problems in household surveys in general and specifically of their impact in applications in urban poverty neighborhoods.

REFERENCES

(1) Jacob S. Siegel, "Completeness of coverage of the Nonwhite Population in the 1960 Census and Current Estimates, and Some Implications." Social Statistics and the City, Report of a Conference Held in Washington, D.C., June 22-23, 1967. David M. Heer (ed.) published by the Joint Center for Urban Studies of the Massachusetts Institute of Technology and Harvard University, 1968.

(2) Ibid., 48-49.

(3) Thirty assignments of approximately 60 households during one week of each month. This is in contrast with the UES pattern, which is weekly assignments of about 10-12 households.

(4) Stanley W. Steinkamp, "The Identification of Effective Interviewers," Journal of the American Statistical Association, 59 (1964) 1167.

(5) George H. Gallup, interview with the New York Times, November 1, 1968.

INTERVIEWER-RESPONDENT INTERACTION IN A HOUSEHOLD INTERVIEW

Kent H. Marquis, Survey Research Center, The University of Michigan

A recently completed study analyzed the verbal interaction that took place between interviewer and respondent when a revised version of the Urban Employment Survey was used. The technique of verbal interaction analysis is a relatively new approach to studying the household survey. It offers some promising insights.

In 1964, with the help of the National Center of Health Statistics, the U.S. Public Health Service, and the U.S. Bureau of the Census, we set out to test a general model of the dynamics of the interview in which respondent demographic characteristics were hypothesized to cause respondent attitudes, motives, levels of knowledge, and ways of perceiving things. It was thought that these psychological variables, expressed in behavior during the interview, were responsible for the accuracy and completeness of household interview data. Five data collections were involved: 1. Health information from the initial interview with the respondent. 2. Interviewer ratings of each initial interview. 3. Counts of interviewer and respondent behaviors during the interview recorded by a third person. 4. Attitude, motive, knowledge and other data about the respondent collected in an additional interview the day after the original interview. 5. A wide range of data about each of the health interviewers obtained by interviewing them at the end of the study.

The results of this study were quite surprising. Measures of interviewer and respondent attitudes, satisfactions, motivations, knowledge, etc., were not correlated with the amount of health information obtained.[1] On the other hand, the greater the rate of behavioral activity of the respondent, the greater the number of items he reported. This held true after correcting as much as possible for factors causing spurious correlations and regardless of whether the behavior was task-oriented or interpersonally oriented.

This study led us to conclude that the major influences on the accuracy or completeness of data obtained in interviews are the actual behavior of the participants and other parameters which have an effect on the immediate situation (e.g., question characteristics). This conclusion is supported by other programs of research relevant to the interview, such as Matarazzo *et al*. (2) on interviewee speech duration, and Rosenthal (3) on experimenter bias.

DESIGN AND PROCEDURES

In the design of the following study the respondent's race and age were the main independent variables and the kind of verbal interaction between interviewer and respondent was the dependent variable. Respondent race (black and white) and respondent age (18-34 or 36-64) were combined in a 2 x 2 analysis of variance design. All respondents were employed males. Four female, middle-class, newly trained white interviewers conducted the interviews. Four separate probability samples, one for each of the respondent groups, were drawn from the population of Detroit, Michigan. One-hundred-and-eighty-one usable interviews were obtained, about 45 in each sample group. Several complicating factors in the design and field procedures caused trouble in obtaining satisfactory response rates in inner-city areas. The response rate was between 50 and 60 per cent.

The questionnaire was a modified version of the 1968 Urban Employment Survey of the U.S. Department of Labor. We retained only the questions applicable to the employed and, of these, we discarded many questions that were asked only of sub-groups in the sample.

The verbal interaction throughout the entire interview was recorded with a small, compact, portable tape recorder. The respondent was given the opportunity to refuse to have the interview recorded and one respondent did refuse. Most recordings were coded directly from the tape. We lost seven interviews (4%) because of inaudible recordings.

The system of behavior coding was adapted from the one used in the study of the health interview. We coded reliably observable behavior such as "asks question," "gives response," "probes," "laughs," etc., rather than inferences like "shows hostility," "is enthusiastic," "appears bored," etc. The complete system of codes, summarized in Table 1, contained 36 categories of verbal behavior. At the end of coder training the inter-coder reliability, computed on a code-by-code basis, was between 75 and 80 per cent.

TABLE 1

SUMMARY OF VERBAL BEHAVIOR CODES

Interviewer

Q	Correct question
∠	Incomplete question
¬	Inappropriate question
X	Incorrect question
=	Repeat question
*	Mistakenly omitted question
N	Question omitted because answer already given
H	Question skipped per instructions
P	Non-directive probe
D	Directive probe
C	Gives clarification
V	Volunteers information

Respondent

R	Adequate answer
W	Inadequate answer
K	Don't know answer
G	Refuses to answer
J	Other answer
E	Elaboration
C	Asks for clarification

Both Interviewer and Respondent

F	Gives feedback
U	Continuing feedback or unsuccessful interruption
T	Repeats answer
A	Irrelevant conversation
S	Gives suggestion
M	Polite behavior
B	Successful interruption
L	Laughs, jokes
Y	Talks to another person
O	Other verbal behavior

RESULTS

Respondents contributed 44 per cent of the verbal behavior in the average interview and interviewers accounted for the remaining 56 per cent. This is in marked contrast to the results of the previous observation study in which respondent behavior constituted a very large majority of the total verbal behavior in the interview.

Interviewer behavior can be classified into three broad areas: question asking, probing, and giving feedback. (See Table 2.) Interviewers are supposed to ask the questions exactly as printed on the questionnaire; according to Table 3, 92 per cent of the questions were asked correctly.

Good interviewer training stresses that non-directive probing is needed. About 22 per cent of all interviewer behavior was devoted to probing of some sort: repeating the question, non-directive probing, directive probing, and repeating answers. For this study, the interviewers behaved the way they were trained. Table 4 shows that about 72 per cent of the probing was non-directive and another 9 per cent consisted of repeating the question, which is another form of non-directive probing. The remaining 19 per cent were directive probes. This was considerably lower than the 42 per cent found in the previous observation study.

TABLE 2

FREQUENCY AND PER CENT OF INTERVIEWER BEHAVIOR, BY CODE CLASSIFICATION

Interviewer Behavior Code Classification	Frequency All Interviews	Per Cent of Total Interview Codes
Asking questions (Q, <, ¬, X)	18,169	37
Probing (=, P, D, T)	11,230	22
Giving feedback (F)	11,498	23
Other (All other codes)	8,948	18
Totals	49,845	100

TABLE 3

FREQUENCY AND PER CENT OF INTERVIEWER QUESTION ASKING BEHAVIOR

Interviewer Behavior	Frequency All Interviews	Per Cent of Total Interview Codes
Questions asked correctly	16,687	92
Incomplete, inappropriate, incorrect questions	1,482	8
Totals	18,169	100

TABLE 4

FREQUENCY AND PER CENT OF 3 TYPES OF PROBING IN 2 STUDIES

	Previous Study	Present Study	
Interviewer Probing*	Per Cent	Frequency	Per Cent
Repeat question	13	638	9
Other non-directive	45	4,938	72
Directive	42	1,269	19
Totals	100	6,845	100

*"Repeats answer" has been left out of the calculations to make the data comparable with the previous SRC observation study.

Table 5 shows that the less directive the probe, the higher the probability of obtaining an adequate answer and the wider the range of behavior which can be elicited. According to theory, non-directive probes should elicit a wide range of behavior. The high rate of adequate answers is less easy to explain. It should be remembered that the interviewer probably does not randomly select the kind of probe that she wants to use and an adequate test of non-directive probing theory requires an experimental design.

The third major classification of interviewer verbal activity is called giving feedback--an interviewer's response to the respondent's behavior ranging from brief signs of approving attention such as "mm-hmm" or "I see" to more elaborate statements like "that's the kind of information we need." Twenty-three per cent of all interviewer behaviors, on the average, fell into this category. This is somewhat disturbing because most programs of interviewer training give very little attention to this aspect of interviewer performance. Two previous studies, which were part of the Survey Research Center Methodological Research Program, anticipated this finding and experimentally varied the interviewer feedback techniques used during the personal interview. They showed that systematic, programmed use of interviewer feedback techniques can increase both the quantity and the accuracy with which health information is reported in the household interview. Data from the present observation study indicate that feedback was used unsystematically. Table 6 shows that the probability that the interviewer gives feedback following an adequate answer is .28.

TABLE 5

PROBABILITIES OF RESPONDENT REACTIONS TO 3 TYPES OF INTERVIEWER PROBES

-Type of Probe-					
Repeat Question		Other Non-directive		Directive	
Respondent Reaction	Probability*	Respondent Reaction	Probability*	Respondent Reaction	Probability*
Adequate answer	.38	Adequate answer	.25	Adequate answer	.19
Inadequate answer	.09	Inadequate answer	.08	Other answer#	.65
Other answer#	.05	Other answer#	.42		
Elaboration	.12	Elaboration	.05		
Asks for clarification	.09	Repeats previous answer	.05		
Repeats previous answer	.05				
Talks to third person	.05				

*Reactions with a probability of less than .05 have been omitted.

#Code J, an answer to anything but a correctly asked question (including a probe) which does not meet the objectives of the question on the questionnaire.

TABLE 6

PROBABILITY OF INTERVIEWER FEEDBACK FOLLOWING RESPONDENT BEHAVIOR BY KIND OF RESPONDENT BEHAVIOR

Kind of Respondent Behavior	Probability that Interviewer Feedback Follows
Adequate answer	.28
Inadequate answer	.24
"Don't know" answer	.18
Refusal to answer	.55
Other answer (Code J)	.34
Elaboration	.30
Repeats answer	.32
Gives suggestion	.33
Other behavior (not classified elsewhere)	.21

The probability that feedback is given following an inadequate answer is .24, approximately the same rate. Even more surprising, if the respondent refuses to answer, the probability that the interviewer thanks him or gives other positive feedback is .55. With this kind of feedback schedule, it is understandable that respondents are sometimes confused about what is expected of them. It also indicates a potential source of interviewer bias.

How do respondents behave? The data in Table 7 indicate that 68 per cent of respondent verbal behavior consists of answers to questions and probes. About 14 per cent consists of elaborations upon answers (furnishing information relevant to the general area of the question but which is not responsive to the exact intent of the question). The remaining 18 per cent of respondent behavior is distributed among 10 different verbal behavior categories. It surprised us to find a 6 per cent rate of asking for question clarification (which is high) and a 2 per cent rate of irrelevant conversation (which is very low).

One of the purposes of this study was to find out whether different respondent groups defined

TABLE 7

FREQUENCY AND PER CENT OF RESPONDENT BEHAVIOR, BY CODE CLASSIFICATION

Respondent Behavior Code Classification	Frequency All Interviews	Per Cent of Total Respondent Codes
Answering questions (R, W, K, G, J, T)	26,455	68
Asking clarification (C)	2,431	6
Elaboration (E)	5,520	14
Irrelevant conversation (A)	850	2
Other (All other codes)	3,645	10
Totals	38,901	100

in terms of age and race would show different patterns of interaction with the female, white, middle-class interviewers. To correct for some minor uncontrolled sources of variation, the results are presented in terms of the proportion of total behavior devoted to each code category. Our sample sizes were not quite large enough for many of the between-group differences to be statistically significant. Therefore, trends in the data will be discussed where they appear to present a consistent and meaningful pattern of results.

In general, respondent age accounts for much more variance in verbal behavior than respondent race, with older respondents showing higher frequencies of almost all categories of behavior than younger respondents and with interviewer behavior following this trend. Elsewhere (Cannell et al, 1) we have referred to this phenomenon as a "general activity level" effect but have made an attempt to remove it in the present data analysis by the "proportion of total behavior" analysis. Even with this "correction," Table 8 data indicate that interviews with older respondents are characterized by a high percentage of most kinds of behavior. Interviews with younger respondents show higher ratios of asking correct questions, giving adequate answers, and asking for and giving clarifications. The interview with older men tends to be more "diffuse" including higher ratios of probes, inadequate answers, interruptions, elaborations, and irrelevancies. Therefore, the interview with younger men is more to the point or "task-oriented," while greater diversity is characteristic of the interviews with older men.

Behavior patterns differ somewhat between interviews with Negro and white respondents. The major characteristic of interviews with blacks (see Table 9) is the predominant focus on task-oriented behavior (behavior directly relevant to communicating questions and answers). Interviewers devote a higher ratio of their behavior to probing and clarification; black respondents give a higher per cent of answers requiring probing and a higher per cent of requests for clarification.

Initially, it was expected that black respondents would show a lower level of cooperation and young blacks would show some overt resistance. However, the pattern observed was one of cooperative, highly-motivated performance.

White respondents have less difficulty with the task than Negro respondents. Evidence for this includes the finding that interviewers probe and repeat questions proportionately less with white respondents and that white respondents give a higher percentage of adequate answers and fewer inadequate responses.

Interviews with white respondents also show a pattern of greater informality typical of everyday conversation, that is, the verbal exchanges include a greater percentage of non-task behaviors like elaborations and interruptions.

In summary, whites performed adequately and engaged in a higher per cent of informal interaction. Blacks had more difficulty but stuck to the task in a highly motivated, businesslike manner.

CONCLUSION

The coded data contain information about specific problems in the interview, including those indicating incorrectly asked questions, questions which the interviewer omitted by mistake, inadequate answers, requests for clarification, etc. Thus, it is possible to use these data to evaluate problems which occur with the questionnaire. We believe that this kind of approach has considerable potential for use in pretests to provide some objective evaluation of the questions. The coded data also disclose information about interviewer behavior such as her propensity to ask questions correctly or incorrectly, her tendency to use directive or non-directive probes, the frequency with which she accepts inadequate answers to specific questions without probing, etc. The examination of the verbal interaction data can provide us with information for feedback to

TABLE 8

AVERAGE PROPORTIONS OF TOTAL INTERVIEWER BEHAVIOR
AND TOTAL RESPONDENT BEHAVIOR BY RESPONDENT AGE

Interviewer Behavior	Average Proportion of Total: Older	Younger
YOUNGER > OLDER		
Correct question	.326	.368
Gives clarification	.062	.066
OLDER > YOUNGER		
Incorrect question	.023	.022
Repeats question	.015	.010
Non-directive probe	.106	.095
Directive probe	.027	.022
Repeats answer	.088	.087
Continuing feedback	.028	.018
Interruptions	.003	.002
Irrelevant conversation	.018	.014
Volunteers information	.023	.021
Laughs	.028	.024
OLDER = YOUNGER		
Incomplete question	.006	.006
Inappropriate question	.002	.002
Feedback	.226	.226
Gives suggestion	.002	.002
Polite behavior	.008	.008
Talks to 3rd person	.010	.010
Total Proportions	1.00	1.00
Number of Interviews	90	91

Respondent Behavior	Average Proportion of Total: Older	Younger
YOUNGER > OLDER		
Adequate answer	.385	.456
Asks for clarification	.054	.066
OLDER > YOUNGER		
Inadequate answer	.086	.077
"J" answer	.174	.158
Repeats answer	.029	.022
Elaborates	.156	.112
Irrelevant conversation	.022	.016
Continuing feedback	.006	.004
Interruptions	.008	.007
Laughs	.028	.026
Talks to 3rd person	.034	.032
OLDER = YOUNGER		
Don't know answer	.006	.006
Refuses to answer	.000	.000
Polite behavior	.002	.002
Feedback	.007	.007
Gives suggestion	.002	.002
Total Proportions	1.00	.99
Number of Interviews	90	91

interviewers about their performance, to evaluate interviewer training, and to test the effectiveness of the different types of interviewer training. Another area which we would like to pursue would use these kinds of data to study how a pattern of interaction becomes established and what kinds of antecedent conditions predict its appearance. The practical implication of this kind of analysis is the ability to predict in what parts of the interview interaction is smooth, at what point fatigue sets in, and at what point the interaction is developed so that more sensitive issues could be introduced.

We made an initial attempt to use the interaction data to make specific evaluations of questions in the urban employment survey. The results obtained were generally very reasonable and useful. Omission problems appeared mostly in questions with skip patterns or other questions which were asked only of sub-samples. Many of the questions which were asked incorrectly contained ambiguous parenthetical statements or awkward syntax. Inadequate answers from respondents appeared frequently to questions requiring respondent recall of past events or to questions where interviewers were unclear about what constituted an acceptable answer. Questions requiring a large number of behaviors to reach an adequate answer were often the same questions with high inadequate answer rates. With further refinements in the diagnostic approach it should be possible to pinpoint exactly which problem exists for each question, as well as the most likely causes of the problem.

The degree to which the findings of this study can be generalized to all survey interviews is of course limited. There is evidence, for example, that an interview on another topic with different kinds of questions, respondents, and interviewers yields different interactive patterns. The purpose of presenting our data is primarily to demonstrate a methodology which shows promise.

TABLE 9

AVERAGE PROPORTIONS OF TOTAL INTERVIEWER BEHAVIOR AND TOTAL RESPONDENT BEHAVIOR BY RESPONDENT RACE

Interviewer Behavior	Average Proportion of Total: White	Negro
WHITE > NEGRO		
Correct question	.358	.334
Volunteers information	.023	.021
Continuing feedback	.025	.022
Interruptions	.003	.002
Laughs	.028	.024
NEGRO > WHITE		
Repeats question	.012	.014
Non-directive probe	.095	.104
Directive probe	.024	.026
Repeats answer	.082	.090
Gives clarification	.062	.065
Feedback	.223	.228
Irrelevant conversation	.014	.018
NEGRO = WHITE		
Incomplete question	.006	.006
Inappropriate question	.002	.002
Incorrect question	.022	.022
Gives suggestion	.002	.002
Polite behavior	.008	.008
Talks to 3rd person	.010	.010
Total Proportions	1.00	1.00
Number of Interviews	90	91

Respondent Behavior	Average Proportion of Total: White	Negro
WHITE > NEGRO		
Adequate answer	.428	.414
Refuses to answer	.001	.000
Elaborations	.140	.128
Feedback	.008	.006
Continuing feedback	.006	.004
Interruptions	.008	.007
Irrelevant conversation	.020	.019
Gives suggestion	.002	.001
NEGRO > WHITE		
Inadequate answer	.076	.084
"J" answer	.154	.176
Asks for clarification	.058	.062
Repeats answer	.024	.028
Laughs	.026	.028
NEGRO = WHITE		
Don't know answer	.006	.006
Polite behavior	.002	.002
Total Proportions	1.00	.99
Number of Interviews	90	91

REFERENCES

(1) Cannell, C.F., Fowler, F.J., and Marquis, K.H. "The influence of interviewer and respondent psychological and behavioral variables on the reporting in household interview." National Center for Health Statistics, Public Health Service, PHS Pub. No. 1000, Ser. 2, No. 26, March 1968.

(2) Matarazzo, J.D. "Studies in interview speech behaviors, in L. Krasner and L.P. Ullman, eds. Research in behavior modification. New York: Holt, Rinehart and Winston, 1965.

(3) Rosenthal, R. Experimenter effects in behavioral research. New York: Appleton-Century-Crofts, 1966.

FOOTNOTE

[1] Several earlier validity studies compared respondents' reported hospitalizations, physician visits, or chronic conditions with medical records. These studies revealed a positive correlation between accuracy of reporting and the amount of chronic and acute conditions reported, probably because in health interviews the problem of underreporting is usually greater than the problem of overreporting. Therefore, the first dependent variable of this observation study was an index of chronic and acute conditions reported for the respondent, correcting for gross variations in real health.

PROBLEMS OF ANALYSIS OF URBAN EMPLOYMENT SURVEY DATA

Howard V. Stambler, U.S. Department of Labor

It would not be far from the truth to say that all new or one-time surveys create unique problems of analysis and interpretation. But in the case of the Urban Employment Survey, which is the focus of our discussion, there may well be a larger than usual number of difficulties. These analytical problems exist even though much of the information being collected in the survey conforms closely to the traditional and familiar Current Population Survey labor force concepts and questions which have gained nearly universal acceptance.

My purpose today is to describe to you some of the major difficulties the BLS is encountering in our analysis of the results and to discuss with you some of our attempts at their resolution. Although only a limited amount of the survey data has been published as yet, there has been an early recognition of the problems that the complete analysis will hold. And, as the BLS begins its examination of the first year's results later this month, these considerations will be well etched in our minds.

To provide a framework for my remarks, let me briefly describe again the purpose and scope of the survey, and how even these create problems in analysis. First of all, the Urban Employment Survey, or UES for short, was designed to examine the employment problems and barriers to meaningful employment of persons in Concentrated Employment Program (CEP) areas--target areas in which the Department of Labor has combined under one administrative structure all its manpower programs in order to concentrate their impact. Its purpose is to provide information for use in understanding the barriers to employment in these and, hopefully, other poverty neighborhoods, so that programs best suited to meet the needs of their residents can be developed. As a result, and this gives rise to our first analytical problem, the UES is less a measurement of the economic and social characteristics of poor people than it is of the situation of all people (only some of whom are poor) who live in specific geographic areas recognized through 1960 Census data as being poverty neighborhoods. And, as we might have expected, these are diverse neighborhoods which include a significant number of persons and families not living in poverty or who do not have serious employment problems.

What this heterogeneity means, is that analyses of the target areas really must be made in three stages or levels. The first level encompasses an investigation of the overall employment situation in the particular poverty area. The next step is the analysis of differences in the situation for the poor and the nonpoor in these areas. Finally, we need to develop as much insight as possible into the particular problems of poor people in these areas.

It is important to note also that these neighborhoods are not necessarily the worst or even the only bad areas in the UES cities. The CEP areas were chosen by the Department of Labor based in large part on the extent of unemployment and poverty in the areas as shown by the 1960 Census. Thus, only part of the total slum area in each city is being surveyed and current employment conditions in these areas may or may not be significantly different from those in other poor neighborhoods in the same cities. In one of the cities--New York--this phenomenon was particularly striking, and tracts outside the CEP area were added to the UES sample.

It is worth noting, too, as was touched on by Mr. Gerson, that the survey is presently being conducted also in the balances of 2 cities, Atlanta and Detroit. Information on these areas will permit comparisons of the employment situation of persons living in the target areas of the two cities with that of the general population outside the target areas. However, this in itself causes problems because just as the six target areas do not have a completely homogeneous population, the balance of the two cities included in the survey are also diverse areas which include other slums, as well as middle- and upper-class neighborhoods. These facts must certainly be kept in mind in making any comparisons between the two sample areas in Atlanta or in Detroit.

Another geographic consideration is the continual population movement into and out of these areas. Slums are increasingly being recognized as "staging areas" for many people who leave them soon after arrival. In addition, as a slum dweller improves his economic status, probably one of his first acts is to move out of the slum. This means that in each succeeding year of the survey we will be measuring changes in the situation in that area, but not necessarily of the people who lived there the previous year. Therefore, any changes in these neighborhoods over time will not necessarily reflect a measurement of the improvement or worsening of the situation for a particular group of people, but rather of the area itself. It will therefore be difficult to detect a change in the condition of particular people by observing changes in the area characteristics.

To compensate for this, we are attempting a longitudinal study in the second year of the survey. Beginning this summer, we are identifying and then following up a subsample of persons who were interviewed during the first year of the survey and have subsequently moved to a new address. This addition to the survey will enable us to find out what happens to both movers and non-movers over time.

In this initial stage of my remarks, I have tried to give you an idea of the geographic composition of the UES areas and of the problems this has created. Let me now turn briefly to one aspect of the sample design that has major analytical implications, with particular reference to statistical reliability.

The weekly sample take in each of the eight survey areas (the six slum areas plus the balance of the cities of Atlanta and Detroit) is approximately 70 households. Thus, over the full year of the survey, about 3,500 households will be interviewed in each area. Since the population size in each CEP area differs considerably, the sampling ratio and thus the estimates of reliability vary somewhat from area to area. For example, in the Atlanta and Los Angeles target areas, approximately one out of every eight households are scheduled for interview each year. At the other extreme, an annual sample of about 1 in 110 persons will be interviewed in the balance of the city of Detroit. Since the same amount of detailed information will be tabulated for each of the areas, analysts and researchers will have to use extreme caution in the interpretation of comparative findings for these areas. Findings which may be statistically significant in one area may be washed out in another area due to the wider range of sampling error. The estimates of reliability are quite different in the UES than in the Current Population Survey, the source of most of our comparative data.

There is a similar problem which arises in relation to the comparison of UES data with information from the CPS--this is the matter of rotation group bias in the CPS.

In the CPS, the monthly estimates are essentially compilations of information obtained from respondents in eight rotation groups. Persons interviewed thus may be in the CPS sample anywhere from the first to the eighth time, since each household is in the sample for 4 consecutive months, drops out for 8 months, and returns to the sample for 4 more months. Long experience with the CPS has shown, however, that the respondents in the first and fifth months of the rotation pattern are more likely to be classified as unemployed than they are in the other months of the survey. By way of contrast, the UES is a weekly survey in which each person is interviewed only once, and the data are accumulated over a given period to provide the necessary reliability. Thus, it is conceivable that all the data from the UES could very well show the same kind of biases as do the first and fifth month households in the CPS.

There are, however, other differences between these 2 surveys which may mitigate such a situation. The length and detail of the UES questionnaire may have some effect on this phenomenon, as could the type of respondent being interviewed in these target areas. Another difference is that the CPS questions for all members of the household are usually answered by one respondent, usually the housewife, whereas in the UES most of the questions are answered by the individual himself (more about this later). In any case, the possible existence of rotation group bias has important implications for analysts, researchers and users of the data who may be making UES-CPS comparisons.

It is worth noting that the UES sampling design will change somewhat during the second year of the survey. Approximately one-half of the UES respondents originally interviewed in one year will be in the sample a second time the following year. The other half of the sample will be comprised of new sample households that will also be interviewed in two consecutive years. Thus, the second year's sample will be comprised of both first and second time respondents with a resulting different rotation group bias effect. This has implications too for UES year-to-year comparisons.

Another source of possible difficulty in the analysis of UES data is the comprehensive and detailed race and ethnic origin information which is being tabulated. Where the sample is adequate and in those cities in which the designation is warranted, data on UES residents will be provided separately for the following groupings--white; Spanish American, both white and nonwhite separately; Mexican American (in Los Angeles and Houston); Puerto Rican (in New York City); and American Negroes. Since the race/ethnic origin composition of the survey areas is quite diverse, it is necessary to be extremely careful in drawing overall conclusions about the data for an entire survey area. Of course, one reason that the particular areas were selected was to enable researchers, policy planners and program administrators to determine to what degree the problems of slum dwellers vary according to their color or national origin.

Although the expanded race-ethnic break will be of great value in examining the situation of different groups in a particular area, it is important to realize that overall data for these particular areas will not be strictly comparable with other area data. For example, a race/ethnic comparison of the situation for whites in the survey area and in the country as a whole would be misleading, since persons of Spanish American birth or parentage make up a very small proportion of the Nation's white population, but about half of the white population in some of the UES areas. Thus, for the analyst and researcher, disaggregation of the data as much as possible will be essential.

Still another area of difficulty which we expect to encounter is in the interpretation and analysis of questions which delve into new and/or sensitive areas. The UES will be attempting to determine if the traditional concepts of "work" and "employment" have the same meaning for poverty residents as they do for most people. It will also attempt to obtain information on illegal or sporadic activity that brings in money but is not generally thought of as work. To do this, the UES will not only use the traditional CPS labor force questions but will also probe further about remunerative activities. For this, we are asking the following questions of all men:

"During the past 12 months did you engage in any kind of activity for which you received money but which you would not normally consider work?"; if "yes", "what was this activity?"; and "Do you have any income from sources other than those previously mentioned, so that you don't need to work?", followed by "what are the other sources of income?", "How long have you been receiving this income?", "How long do you expect it to continue?", "Is it enough for your needs?", and "Do you expect to go to work at a regular job?". There is, of course, the possibility that the questions will not be properly understood, but even when they are understood, what degree of reluctance and evasiveness will we encounter from respondents? What degree of credibility can we attach to the responses? Despite our wariness about these questions, we feel that this probing may provide some valuable insights into those areas.

Other, more traditional types of questions which may be viewed as "sensitive" ones are those on earnings and income. The UES is collecting information on weekly earnings for those employed in the week prior to the survey and on total personal and family income over the past year. Based on past survey experience, we know that there is often a reluctance on the part of respondents to divulge sources and amount of income. In addition, difficulties in recollection over long periods cast doubt on the accuracy of income data for an entire year. But even more important, how will these particular target groups respond to these questions? Will poor people be more reluctant or less reluctant to discuss levels and types of earnings and income, especially welfare payments or illicit income. These facts also have to be taken account of in our analysis.

There is a separate aspect of the survey which may pose some difficulty to analysts and researchers, since it represents quite a departure from the usual objective and factual socioeconomic data that the Bureau of Labor Statistics has traditionally been concerned with. The new survey delves into the very subjective area of attitudes about jobs and about life in general, motivations, aspirations, and opinions. In addition to being a departure from the usual work of the Bureau, further difficulty in analysis will be encountered due to the fact that there are little comparable data for persons not in poverty neighborhoods by which to gauge survey results.

Two different types of information will be available from these subjective questions. The first is on what might broadly be called "job satisfaction." A series of questions will probe the attitude of poverty residents toward their jobs, determining what they like or dislike about the job, their commitment to work, and whether the job plays a meaningful role in their lives. These will be correlated with many characteristics, such as occupation, education, and income to develop some insights into whether attitudes of poverty residents are significantly affected by these characteristics.

The second general area relates to discrimination as perceived by poverty residents. This too will be analyzed in relation to many of the job characteristics and attitudes of poverty residents. Although responsibility for this area of work will be in the hands of social scientists with the proper background and training, we nonetheless expect to encounter many problems in the proper analysis and interpretation of the new data.

I have already discussed problems of comparability in relation to some of the unique characteristics and concepts of the Urban Employment Survey. But the UES data will have other problems of comparability, even where it uses traditional labor force concepts and measures. The major difficulties in the comparison of seemingly similar measures, such as from the UES and CPS, arise mainly from the different time periods covered. As I indicated, the UES is an accumulated sample, and the information will be analyzed and published after a full year's collection. This means that annual data on the current employment status of UES residents will actually be an accumulation of the situation for 52 different weeks, as opposed to the CPS, in which an annual average is an arithmetic average of the 12 monthly observations covering the same week of each month (the week containing the 12th).

Similarly, other data for the annual periods covered in the UES will be quite different than in the CPS. Data on annual income and annual work experience in the CPS are collected in February and March each year, and always refer to the calendar year prior to the survey. In the UES, however, data for the previous year's work experience of income will, in essence, relate to a sliding reference period covering the 12 months immediately prior to the week in which the interview was held.

For income and other questions, this also has important recall implications. It is probably easier to recall income for a calendar year (particularly when the questions are asked around income tax time as in the CPS) than it is to recall income for a June-to-June or a September-to-September period. Similarly, there is some doubt as to whether the income response covering the past 52 weeks may not in reality cover merely the previous calendar or income tax year. This sliding reference period also affects comparability with other CPS data usually obtained from supplementary questions to the monthly survey (such as educational attainment information) which always relate to a single month. In the UES, they will be accumulations of weekly responses covering a particular period. Differences like these in periods covered in the two surveys also raise questions of both seasonality and differences in economic climate at varying periods of time, items which must be considered in the analysis of UES data.

There are still other problems and difficulties mentioned by the previous speaker, which I will note only in passing. These are the serious problems of underenumeration, the lack of population controls for the slum areas, and non-response. But there is one last problem in the UES that I would like to mention in closing, one posed by the mode of response. As I indicated earlier, in the UES the household respondent answers only current employment status questions for other members of the household. However, the bulk of the questions, such as on last job held for persons not in the labor force and on work experience, are answered only by the individual himself. In addition, if the person indicates his current employment status to be something different than that reported by the household respondent, answers are changed to reflect the individual's responses. In the CPS, on the other hand, the housewife or whoever is at home at the time of the survey is the person who usually answers the questions on labor force status and work experience. All of these differences must be taken into account by users of the data.

Let me now conclude with one general comment which is undoubtedly needed to place these analytical problems in perspective. I certainly do not want to leave the impression that the data will be unusable or that the survey results will not be meaningful; quite the contrary. Despite the acknowledged limitations of the data, we feel that the information from this survey will be extremely valuable to researchers, program planners, policy makers and others. It can not only provide new insights into barriers to employment in poverty areas but can also help to uncover problems about which more should be known. Nevertheless, the interpretation and analysis of the UES findings will present major challenges for both the Bureau of Labor Statistics and other users; caution and restraint will have to be essential tools for all those who use the data.

DISCUSSION

Curtis C. Aller
San Francisco State College

Candor compels me to admit, as I did to those who invited me, that as a non-statistician I could not provide the critique these papers so richly deserve. Fortunately, a discussant is allowed to pick his targets. In my case I plan to speak briefly from the perspective of one of those who pushed vigorously for this endeavor and as a prospective user of the results.

From this perspective, then, the questions that are of concern to me are: How far has the BLS been able to go towards accomplishing our original objectives? How good a job (in a technical sense) is being done? Has the effort been worthwhile?

Mr. Stambler has commented quite adequately on the objectives as they are seen by the Bureau but let me elaborate a bit further. The Secretary of Labor, Willard Wirtz, was fond of saying that our usual local employment figures were not relevant and in fact tended to be misleading. We concluded, therefore, that our minimum objective had to be the development of some valid labor force measures for the poverty areas of our large cities. This, clearly the UES is now doing and as time passes the data yield will become even richer.

Some hoped we could also develop a tool that would measure the impact of our various poverty-manpower programs in these areas and thereby determine our progress over time. Mr. Stambler has noted one very real difficulty here as the UES measures areas that serve as staging centers for the poor. The associated effort to measure outflow of people from these areas may give us some insights into program effects. I suspect, though, we will find it necessary to couple these studies with similar studies designed to measure the in-flows of people into these areas as well. This information will turn out to be valuable to many users and particularly so, I think, to manpower program planners. Program evaluation involves other and extremely complex issues and I am convinced that the UES together with population flow studies will not provide anything like the quick and conclusive evaluative tool some had expected.

Still others (and in this category I place myself) viewed the UES as an exploratory research effort. We were determined to avoid the development of another statistical series as the sole outcome. Therefore we urged that some of the resources be devoted to special studies at the cost, I might add, of coverage of additional cities. The undercount phenomena so strikingly demonstrated in Gerson's paper was one obvious example. Another would be an effort to probe more deeply into the attitudinal arena and more specifically to undertake an exploration of the meaning of work. A third dimension involved income questions. Thus we thought we needed to know more about the actual income levels and sources of income so that we could begin to test more accurately the transfer costs and requirements for moving from one way of life to that contemplated by our manpower programs. In this connection we already had considerable information indicating that the social minimum wage for many in these areas was well above that being achieved by our manpower training efforts. The facts are there but the reasons are by no means clear.

In these respects the concern was in finding out if we were asking the right kind of questions --that is, were our conventional labor force concepts the meaningful ones and therefore could continue to serve as the appropriate foundation for the superstructure of programs we had been erecting in recent years.

On this dimension I think we can conclude the results have been promising. The BLS has been prepared to break out of its historical mold. The record demonstrates that some innovative special studies have been designed and are underway.

The fact that the BLS is now moving ahead on a pioneer effort to use the technique of participant observers in combination with the UES is a striking example of its capacity for change. William F. Whyte, the chairman of the advisory committee on UES, has described this venture as a great scientific challenge. Put simply the challenge is to our ability to discover how we can go from impressionistic data to the hard without losing reality.

This addition to the UES is a risky venture in the sense that it may produce many new and significant insights or it could be an utter failure, producing some interesting case studies but nothing of value for comparative purposes or leads for new or restated questions for the UES.

I find it comforting that the Department of Labor remains willing to embark on such a speculative adventure and, particularly so, in view of the wholesale change of leadership that has come with a new administration.

On a more technical level I would conclude that both Gerson and Stambler between them have covered the important problems. To Stambler caution and restraint are the bywords. I've observed this to be the usual stance of the statistician who often contemplates with consternation and horror the liberties taken with his data by politicians and lesser policy makers. But in this case I think the caution is genuinely merited. I'm confident the BLS will continue to stress restraint with their usual skill.

On the other hand I am comforted by Gerson's conclusion that in some areas as, for example, the rotation bias and the comparability of the CPS and UES that these can in time be measured or at least estimated. I'm sure he is right and users will be able to rely on the resultant information with increasing confidence.

Finally, I'm impressed with the Marquis paper as an example of careful scientific work. First, he builds on previous research so as to concentrate on the major influential factor suggested by these earlier studies--the interaction of respondent and interviewer. I'm impressed by the ability--to me a surprisingly high 50 to 60%--to secure respondent participation in taped proceedings. The method seems to offer a means for improving questionnaire design beyond that now possible with usual pre-test methods as well as a device for training and developing interviewers.

I have only one small question to raise. Does the difference of White-Negro respondents, the informality of everyday conversation for the Whites versus the task-orientation of the Negroes relate in part to the exclusive use of White females as interviewers?

I can only conclude that we have been presented with a set of excellent papers. I have nothing but admiration to express for the high level of competence reflected therein and by inference for the quality of the UES.

Walter B. Miller, Joint Center for Urban Studies of MIT and Harvard

The task of obtaining useful information as to unemployment and related problems in urban slum districts presents both unusual opportunities and unusual difficulties. Earle Gerson and Howard Stambler, in their papers dealing with problems of interviewing, methodology, and analysis of the Urban Employment Survey of the United States Department of Labor and the Census Bureau, have chosen to stress the difficulties; I would like to take the opportunity, in my capacity as a discussant, to point to the not inconsiderable merits of this enterprise. It has taken a good deal of courage and imagination to depart from the well-established and relatively comfortable traditions of Census Bureau data-gathering methods in order to tackle an extremely important area of information. It is my feeling that the Urban Employment Survey breaks new ground in the field of government-sponsored information gathering, and that the difficulties of the enterprise are in direct proportion to its significance. Marquis' paper shows a similarly imaginative approach to some very fundamental methodological problems of the survey-research method.

Because I so applaud the spirit of this enterprise as well as its execution to date, I am most reluctant to be critical of it. However, in order to conform to the traditional role of the cultural anthropologist, and particularly, in this case, a cultural anthropologist with very limited statistical sophistication, I feel obliged to indicate certain limited aspects of these studies which appear to me to have received insufficient systematic attention, and where the application of somewhat more attention in the future might result in significant improvements in the character and usefulness of the research findings.

Of the two classic statistical issues of "reliability" and "validity", it is my feeling that both Gerson and Stambler show relatively more concern with the former. The Marquis paper does go to the problem of validity, but on a rather restricted level. I would like to address briefly and in more general terms, some problems of validity as they apply in studies of this kind.

Marquis, in referring to the results of his Health Statistics surveys and related studies, cites this conclusion: "...the major sources of influence on the accuracy or completeness of data obtained in the interviews are to be found in the actual behavior of the participants and in other parameters which have an effect on the immediate situation." I would like to dispute this conclusion, and suggest that considerably more important sources of influence on the accuracy of information relate to certain larger subcultural systems which figure directly in this kind of data-gathering enterprise. Two of these are of particular relevance here; the subculture of the respondent populations, and the subcultural context of the survey-research method in general.

I will use the term "subculture" to refer to those sets of conceptions, perceptions, and definitions of appropriate practice maintained by designated categories of persons, such as males, adolescents, and urbanites, by virtue of their affiliation with that category. Of particular relevance here are subcultures associated with different social status levels. I will refer, in a very gross fashion, to "middle class", "working class", and "lower class" subcultures, although a good deal of additional refinement with respect to differentiation of levels is obviously possible, and utilized elsewhere.

Survey-research methods in general, including the particular subtypes of that method employed in the federal census, were developed primarily in connection with middle-class and/or working class populations. The philosophy behind the method as well as its methodological feasibility rests on the presumed existence of a large set of conventional attitudes and customary behavioral practices characteristic of such populations. Among these are relatively stable residence, certain types and levels of communication patterns, a motivation to provide accurate answers, a sense of obligation to aid the purposes of public data-gathering enterprises, and many others. Populations at the lowest social status levels--and in particular the urban low-skilled laboring class--maintain an equally conventional set of life patterns which differ in important respects, from those of the middle and/or working class. Insofar as there are differences between higher and lower status populations which affect the probability of obtaining valid information through survey research methods, it would appear that this probability decreases as one moves down the social scale. In many cases these differences are not radical ones, but they are systematic, not random, and of sufficient magnitude as to have systematic effects on results.

The research in urban employment now under way provides an excellent opportunity to explore some of the effects of differential social status on the validity of survey results--an opportunity which has not been sufficiently exploited to date. For example, it is most significant that Marquis makes his primary analytic distinctions on the basis of age and race, and neglects social status. I would suspect that the use of a sufficiently refined method of discriminating social-status levels (one based, for example, on years of education, an item in the present questionnaire) would show substantially better discriminations than race. I say the opportunity is an excellent one primarily because of one of the major empirical findings of the employment studies to date; namely, the unexpectedly high degree of social-status heterogeneity in the Concentrated Employment Program areas. It would not be too difficult, on the basis of data now available, to distinguish three, four, or more status levels within the lower class populations of these communities, and use these distinctions as a major dimension in the analysis of variation.

I would like to note briefly some influences of social-status subcultures on the validity of questionnaire and/or interview responses with respect to two orders of information: the first, essentially "factual" information; and the second, attitudinal. Stambler's formulation reflects this kind of distinction in his citation of "objective and factual" versus "subjective" data (a further issue, the degree of "objectivity" of the most "objective" information, cannot be treated here). Subcultural factors affect both orders of information. One can distinguish two kinds of influence; the first might be called "inadvertent misrepresentation" and the second "motivated misrepresentation". Included under the first are familiar factors such as faculty memory (phrased more impressively as "inaccurate informational retrieval"), and cognitive misunderstanding of questions (Marquis' paper treats this subject), based either on inadequate communication a "true" incapacity to grasp the intent of the question, or both.

It is also possible to distinguish two kinds of "motivated misrepresentation"--"conscious" and "unconscious". Conscious motives may involve the desire to "stay out of trouble", involving, for example, information concerning illegal sources of income (numbers, prostitution, theft) or unreported income from a job held by public welfare recipients. Less conscious motives may entail misrepresentation in an effort to create a "good impression" in the eyes of the interviewer--representing one's practices, by subtle shading or more gross distortion, in such a way as to conform to the respondent's conception of what the interviewer's conception of "proper" or acceptable forms of behavior might be. The influences of more- versus less-conscious bases of misrepresentation are difficult to separate out for particular responses, but may be distinguished on a conceptual level.

One way of approaching these very difficult problems respecting the validity of information collected among low-status populations would be to select a limited number of subsamples of the Employment Survey study populations, and subject these to highly intensive research--based not on the standard survey approach of one, two, or three interview contacts of one-half to one hour each, but rather on the long-term, continuing, direct-contact methods developed in anthropological field study, and often referred to as the "participant observation" approach. Such research would center on exactly the same informational areas as those covered in the Employment Survey questionnaire, but obtain this information through very different means. If one could arrange a coverage of even 5 or 10% of the total Employment survey sample by this method, much of the uncertainty as to the validity (although not necessarily the reliability) of the questionnaire responses could be reduced.

I cannot go into detail as to the specifics of this method, which has both assets and liabilities for present purposes, other than to say that it has been developed in some detail, and has shown excellent results under circumstances where survey methods would be quite inappropriate--for example, the study of urban adolescent corner gangs--many of whose most important activities involve illegal practices which probably would not be reported to an unfamiliar interviewer, or if reported, reported inaccurately. Such an enterprise is in fact currently being planned under the auspices of the Labor Department, and specifics might perhaps be reported at some future session. One of the major features of this method is the opportunity it affords to record relatively independent bodies of data with respect to two orders of information: "expressed sentiments" (statements involving attitudes, intentions, and/or verbal characterizations of current situations) and "observed practice" (what individuals are actually observed to do with respect to areas to which their verbal statements refer). An informant might tell an interviewer, for example, that he is "actively seeking" employment, whereas continued observation of his behavior over three or four months might reveal his actual search to be quite casual or even non-existent.

Since the papers by both Stambler and Gerson have stressed methodological deficiencies in data collected in the present studies, and I have indicated still others, I would like to conclude by posing, but in no sense resolving, what can be seen as a fundamental issue with respect to the informational product of such studies. In areas of high policy relevance where very little reliable data are available, are the interests of effective policy formulation better served by the availability of information incorporating known and systematic sources of bias, or by continuing to base policy on those more impressionistic and less systematic forms of information which are available? I am by no means convinced that the former alternative is the preferable one, but see this question as one which each concerned individual will resolve according to his own values and special interests.

IV

SURVEY METHODOLOGY: BIAS AND ITS REDUCTION

Chairman, JACOB J. FELDMAN, Harvard University

APPLICATION OF THE RANDOMIZED RESPONSE TECHNIQUE IN OBTAINING QUANTITATIVE DATA

Bernard G. Greenberg, University of North Carolina at Chapel Hill
James R. Abernathy, University of North Carolina at Chapel Hill
Daniel G. Horvitz, Research Triangle Institute

1. INTRODUCTION

Refusal to respond and the deliberate giving of false information are known to be two principal sources of non-sampling error that can bias sample estimates in surveys involving human populations. Recognizing that such evasive answers are more frequent when respondents are queried about sensitive or embarrassing matters, Warner [1] developed an interviewing procedure designed to reduce or eliminate this bias. He called the technique "randomized response" because the participant in the survey responds to a question which he selects at random from one or more questions, and does so in such a manner that the interviewer does not know which question is being answered.

Warner considered the case where a proportion π of the population (say Group A) possessed some sensitive characteristic while the remainder of the population did not possess the characteristic. The objective was to evaluate π. With the aid of a randomizing device, the respondent selects one of the following statements by chance,

I am a member of Group A.
I am not a member of Group A.

and answers "Yes" or "No" to whichever one of the two statements is selected. The interviewer does not know to which question the reply refers and, indeed, does not want to know.

The rationale underlying the randomized response procedure is that, since the respondent can answer the question without revealing his personal situation, potential embarrassment and stigma have been removed and, in the process, the primary reason for refusals to respond and evasive answers has vanished. If respondents are fully convinced that there is no need to conceal or falsify the facts, it follows that cooperation of the respondents and validity of their responses will be improved.

Abul-Ela et al [2] extended Warner's model to the trichotomous case designed to estimate the proportions of three related, mutually exclusive groups, one or two of which possessed a sensitive characteristic. The model was further extended to estimate any j proportions ($j>3$) when all the j group characteristics are mutually exclusive, with at least one and at most $j-1$ of them sensitive. The reason behind this extension was to provide theory for the multichotomous situation which is often found in nature and which is, perhaps, more realistic than the dichotomy.

Another recent development in the randomized response technique is the unrelated question model. The Warner technique is concerned with two questions (or statements) both of which are related to the sensitive characteristic. The unrelated model is predicated on the assumption that the confidence of the respondents toward the anonymity of the technique would be increased if two unrelated questions were used, one pertaining to the sensitive characteristic and the other to a non-sensitive, innocuous condition. Abul-Ela described this model in [3], Horvitz et al [4] elaborated upon it further and presented results from two field studies which utilized two unrelated questions. Greenberg et al [5] studied theoretical aspects of the unrelated question technique and compared it with the original Warner model. They concluded that the unrelated question technique was more efficient than the original model and recommended it for general use.

There are several other ramifications of the randomized response technique currently being explored by the investigators cited above, and others, such as two trials per respondent and having the innocuous question built into the design. Up to the present time, practically all the research in the field of randomized response has been concerned with refining the technique for use with questions of a qualitative nature requiring only a "Yes" or "No" response. The technique need not be restricted to nominal scale data. The objective of this paper is to present the results of obtaining quantitative data on a highly sensitive subject (abortion) and on one considered to be only moderately sensitive (earnings of head of household). Emphasis is on derivation of estimation procedures relative to mean and variance of the distributions. The data were collected in the North Carolina Abortion Survey described in [6] and briefly reviewed here.

2. COLLECTION OF DATA

The randomized response device was a sealed, transparent, plastic box approximately 4" long, 3" wide, and 1" deep. The device used in the abortion trial had two questions printed on the box lid:

1. How many abortions have you had during your lifetime?

2. If a woman has to work full time to make a living, how many children do you think she should have?

The first question had a small red ball and the second question had a blue ball attached in the space occupied by the number "1" and "2" in the questions shown above. Inside the box there were 35 red and 15 blue balls. The respondent was asked to shake the box of freely moving balls

This research was supported under Research Grants HD03461-01 and HD03441-01 from the National Institute of Child Health and Human Development.

thoroughly, and to tip the box allowing one of the balls to mount a built-in runway and appear in a "window" which was clearly visible to the respondent. The color of the ball which appeared in the window determined which of the two questions the respondent was supposed to answer. If a red ball appeared, she answered the question which had a red ball in front of it (abortion); if a blue ball appeared, she answered the question which had a blue ball in front of it (number of children).

The interviewer was some distance away from the respondent and, of course, did not know which question had been selected by the device. The respondent's reply was simply a number, without specifying to which question the answer referred. To prevent identification the box was shaken again before returning it to the interviewer.

Two independent, non-overlapping stratified cluster samples of women 31 years of age or older were drawn from the population. Total sample size was optimally allocated to the two samples as discussed in [5] for the purpose of producing minimum variance estimates of parameters other than those considered here. The two sample sizes, n_1 and n_2, consisted of 623 and 287 women, respectively.

The procedure used in the trial for earned income was identical to that described for the abortion study. The respondents were not the same group of women answering the abortion set of questions. They represented an independent sample of 1628 women 18 years old or over from the same metropolitan areas, with n_1 and n_2 equal to 1040 and 588 women, respectively. The income questions were as follows:

1. About how much money in dollars did the head of this household earn last year?

2. About how much money in dollars do you think the average head of a household of your size earns in a year?

The response to these questions was a sum of money.

3. QUANTITATIVE RANDOMIZED RESPONSE MODEL

In the quantitative model utilizing two questions, the overall distribution of responses is comprised of numerical answers to both questions, and are indistinguishable. This distribution is, therefore, a mixture of two "pure" distributions which must be statistically separated in order to provide meaningful estimates of the parameters of interest, the population mean of both the sensitive and non-sensitive distributions, μ_A and μ_Y, respectively, and variances, σ_A^2 and σ_Y^2.

Assuming two independent, non-overlapping samples of sizes n_1 and n_2, let

P_i = the probability that the sensitive question is selected by the respondent in sample i. (i=1,2)

$1-P_i$ = the probability that the non-sensitive question is selected by the respondent in sample i. (i=1,2)

Z_{ij} = response from individual j in sample i. (i=1,2), (j=1,2,...,n_i)

f(A) = probability density function associated with the sensitive question.

f(Y) = probability density function associated with the non-sensitive question and similar to f(A) except for location parameter.

$\bar{A}$ = sample estimate of the mean of the sensitive distribution.

$\bar{Y}$ = sample estimate of the mean of the non-sensitive distribution.

The probability density functions of Z_{1j} and Z_{2j} may be written as:

$$\text{Sample 1:}\quad f(Z_{1j})dZ = P_1 f(A_{1j})dA + (1-P_1)f(Y_{1j})dY \qquad j=1,2,\dots,n_1$$

$$\text{Sample 2:}\quad f(Z_{2j})dZ = P_2 f(A_{2j})dA + (1-P_2)f(Y_{2j})dY \qquad j=1,2,\dots,n_2$$

Then, by looking at first moments and taking expected values, we get

$$E(Z_1) = \int Z_1 f(Z_1)dZ = P_1\int Af(A)dA + (1-P_1)\int Yf(Y)dY$$
$$E(Z_2) = \int Z_2 f(Z_2)dZ = P_2\int Af(A)dA + (1-P_2)\int Yf(Y)dY \tag{1}$$

or

$$\mu_{Z_1} = P_1\mu_A + (1-P_1)\mu_Y$$

$$\mu_{Z_2} = P_2\mu_A + (1-P_2)\mu_Y$$

Hence, we may write

$$\mu_A = \frac{(1-P_2)\mu_{Z_1} - (1-P_1)\mu_{Z_2}}{P_1 - P_2}$$

$$\mu_Y = \frac{P_2\mu_{Z_1} - P_1\mu_{Z_2}}{P_2 - P_1}$$

By substituting the estimated responses, $\bar{Z}_1$ and $\bar{Z}_2$ respectively for μ_{Z_1} and μ_{Z_2}, unbiased estimates, say $\bar{A}$ and $\bar{Y}$, are obtained for μ_A and μ_Y. Thus,

$$\bar{A} = \frac{(1-P_2)\bar{Z}_1 - (1-P_1)\bar{Z}_2}{P_1 - P_2}$$
$$\bar{Y} = \frac{P_2\bar{Z}_1 - P_1\bar{Z}_2}{P_2 - P_1} \tag{2}$$

The estimated variances of these unbiased estimates of μ_A and μ_Y are straightforward, viz.

$$\hat{Var}(\bar{A}) = \frac{1}{(P_1-P_2)^2}\left\{(1-P_2)^2 Var(\bar{Z}_1)+(1-P_1)^2 Var(\bar{Z}_2)\right\}$$

and (3)

$$\hat{Var}(\bar{Y}) = \frac{1}{(P_2-P_1)^2}\left\{P_2^2\ Var(\bar{Z}_1) + P_1^2\ Var(\bar{Z}_2)\right\}$$

Maximum likelihood estimates of μ_A and μ_Y can not be derived without further specification of f(A) and f(Y) in the following likelihood function, viz.

$$L = \prod_{j=1}^{n_1} [P_1 f(A_{1j})dA + (1-P_1)f(Y_{1j})dY] \cdot \prod_{j=1}^{n_2} [P_2 f(A_{2j})dA + (1-P_2)f(Y_{2j})dY]$$

Using either the normal and Poisson distributions simultaneously for f(A) and f(Y) in the foregoing equation leads to expressions which are not easily tractable. Therefore, an investigation of the asymptotic relative efficiency of $\bar{A}$ and $\bar{Y}$ compared to their respective maximum likelihood estimates is now in process. The lower bounds of σ_A^2 and σ_Y^2 may be obtained by examining the negative of the expected value of the second derivative of ln L with respect to the parameters. Unfortunately, these equations are also cumbersome and individual specific cases are being checked on the computer.

4. NUMERICAL EXAMPLES

The North Carolina Abortion Study was carried out for the dual purpose of evaluating the randomized response technique as a means of collecting data on sensitive matters and to estimate induced abortion rates in an urban setting. Probability samples of adult women were selected from five metropolitan areas of North Carolina as already indicated. The population of these cities ranged from 100,000 to 250,000 persons. We had four different samples and each sample was selected for a different purpose. Two of the four samples were used to obtain quantitative data on abortion and income.

Abortion and number of children. Data from the North Carolina Abortion Survey are shown in Table 1. It was desired to estimate the mean number of abortions over a lifetime (μ_A) by color, and the mean number of children a woman should have if she had to work full time to make a living (μ_Y), by color. Estimated variances of both estimates were also desired. In this survey the probability, P, that the sensitive question (abortion) was selected by the respondent - the ratio of red balls to total balls in the plastic box - was 0.7 in Sample 1 and 0.3 in Sample 2.

The data in Table 1 are followed by estimates of the parameters. Variances shown were computed under the assumption of a simple random sample study design.

Table 1. Frequency Distribution of Responses to Abortion Set of Questions by Color and Sample, North Carolina Abortion Survey, 1968

Z Values	Number of Respondents			
	Sample 1		Sample 2	
	White	Nonwhite	White	Nonwhite
0	304	130	114	56
1	14	22	10	8
2	56	39	30	17
3	10	9	6	2
4	7	8	1	5
5	2	-	1	2
6	-	-	-	1
Unknown	13	9	25	9
Total	406	217	187	100
$\bar{Z}$ (excluding unknowns)	.494	.764	.599	.923

Estimates of parameters:

Color	$\bar{A}$	$\sqrt{\hat{Var}(\bar{A})}$	$\bar{Y}$	$\sqrt{\hat{Var}(\bar{Y})}$
White	.415	.107	.678	.145
Nonwhite	.645	.177	1.042	.268

It is clear that the estimated values of $\bar{A}$ and $\bar{Y}$, by color, appear reasonable. For instance, results in the field study by Abernathy et al [6] indicated that abortion was more common among nonwhite women.

It is also clear from observation of the Z values that they do not appear to be a sample from a linear combination of two distributions having the same probability density function. The excess of 0's and 2's would suggest some preferential grouping by respondents. If true, further work in investigating the appropriate analysis will have to be carried out.

Income. Estimates of mean income of heads of households were computed in the same manner. In this experiment the probability that the respondent selected the first question was 0.64 in Sample 1 and 0.36 in Sample 2. Mean income of the "head of this household" ($\bar{A}$) and mean income of the "average head of a household of your size" ($\bar{Y}$) are shown in Table 2 by color. Estimated variances assuming a simple random sample design are also presented.

Table 2. Estimates of Means and Variances of Income Set of Questions by Color, North Carolina Abortion Survey, 1968.

Color	Estimates (in dollars)			
	$\bar{A}$	$\sqrt{\hat{Var}(\bar{A})}$	$\bar{Y}$	$\sqrt{\hat{Var}(\bar{Y})}$
White	6690	418	8001	446
Nonwhite	3370	445	4530	588

These estimates of income also appear reasonable for the particular population under study. Earnings of nonwhites are considerably less than for whites as would be expected. Furthermore, both races visualize the economic standing of the "average household head" as being superior to their own. This, too, might be expected.

5. SUMMARY

The development of the randomized response technique from its inception as a survey technique to eliminate evasive answer bias in sensitive questions of a dichotomous nature to the present state of the art is examined. An attempt is made to extend the method to questions which elicit quantitative rather than qualitative responses.

The theory underlying the quantitative application of the randomized response procedure is presented and applied to data collected for that purpose in a recent abortion survey in North Carolina. Estimates and variances of the mean number of abortions obtained over a lifetime in an urban population of women, and mean income of heads of households are reported.

REFERENCES

[1] Warner, S. L., "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias," Journal of American Statistical Association, 60, (1965), 68-69.

[2] Abul-Ela, Abdel-Latif A., Greenberg, Bernard G., Horvitz, Daniel G., "A Multi-Proportions Randomized Response Model," Journal of American Statistical Association, 62, (1967), 990-1008.

[3] Abul-Ela, Abdel-Latif A., "Randomized Response Models for Sample Surveys on Human Population," Unpublished Ph.D. Thesis, University of North Carolina, Chapel Hill, 1966.

[4] Horvitz, Daniel G., Shah, B. V., Simmons, Walt R., "The Unrelated Question Randomized Response Model," Proceedings of Social Statistics Section, American Statistical Association, 1967.

[5] Greenberg, Bernard G., Abul-Ela, Abdel-Latif A., Simmons, Walt R., Horvitz, Daniel G., "The Unrelated Question Randomized Response Model: Theoretical Framework," Journal of American Statistical Association, 64, (1969), 520-539.

[6] Abernathy, James R., Greenberg, Bernard G., Horvitz, Daniel G., "Estimates of Induced Abortion in Urban North Carolina," Paper to be published in Demography, February, 1970.

THE POTENTIAL SYSTEMATIC BEHAVIOUR OF SOME PANEL SURVEY ESTIMATES

W. H. Williams and C. L. Mallows, Bell Telephone Laboratories, Incorporated

ABSTRACT

Panel surveys involved repeated observations made on the same sample of population units. In some of these studies systematic biases have appeared; specifically, if R_i is an estimate made from panels appearing in the sample for the i^{th} time, then it has been observed that these estimates sometimes vary systematically with i. It is tempting to interpret these phenomena as being due to conditioning of the panel of respondents.

In this paper it is shown that these systematic changes can be a result of the characteristics of the response probabilities. This is true whether the estimates are computed only for those individuals who are identical from one interview time to the next, or on all available persons, or on those persons appearing only once.

A discussion of similar results for complete followup surveys, and the description of a simple "adjusted" estimation procedure are also included in the paper. Some numerical examples are presented to show that in some very innocuout situations the potential biases can be very large.

I. INTRODUCTION

Panel surveys involve the periodic observation of the same population units. In some such studies, systematic biases have appeared which do not appear to have been fully explained. For example, in the Current Population Survey conducted by the U.S. Bureau of the Census [7], the unemployment rate of those persons interviewed for the first time appears to be considerably higher than on subsequent interviews, see Waksberg and Pearl [8] and the Report of the President's Commitee to Appraise Employment and Unemployment Statistics [9]. One of the main purposes of this paper is to show that, under certain simple hypotheses each of several natural estimates made at two different interview times, may exhibit systematic changes simply as a result of sampling characteristics.

Panel surveys have many variations. Sometimes a selected sample is followed up as completely as possible at each of the subsequent observation times. On the other hand rotation designs involve the month-to-month (say) retention of some sampling units and the replacement of others. These later designs can be used to produce good estimates at specific points in time as well as good estimates of change. The details of rotation sampling and panel studies in general will not be described in this paper because there is a large literature on the subject, see for example Hansen, Hurwitz and Madow [1], Cochran [2], Patterson [3], Eckler [4], Rao and Graham [5], and Kish [6].

Generally, if R_i is the estimate made from those rotation groups which are appearing in the sample for the i^{th} time, then it has been sometimes observed that there are significant and systematic changes in these estimates when considered as a function of i. A study of this problem by one of the authors lead to a possible explanation of these phenomena; see Williams [10]. In that paper it was shown that if, (1) the probability of nonresponse for a selected unit is monotonically related to the characteristic under study, and if (2) this same probability changes from one observation time to the next, then systematic changes in the expectation of the estimator must occur.

These two hypotheses are very reasonable ones. It is well known that nonresponse is sometimes related to the characteristic being measured. For example, families with no children are much more likely to be missed than families with children. There is a large literature on this type of behavior, see for example Kish [6]. Also there can be little doubt that the probabilities change from one interview time to the next. This is often clearly revealed by a changing nonresponse rate, which can change only if the nonresponse probabilities change; and after all it is the continuing goal of every survey manager to achieve a higher response rate.

The question remains, however, as to whether any of the observed systematic changes could be caused by other factors. For example, is it true that people tend to answer questions about their economic status differently in a first interview than in a second? To examine this question it is tempting to construct estimates which are based only upon those individuals who appear in the survey both at T_1 and T_2. The argument is that systematic changes in the estimate for this matched set must be the result of factors other than purely statistical ones. This conclusion is false. The reason is that systematic changes in the probability of nonresponse may be ensuring that the matched set of individuals is not representative.

The same difficulty can arise in <u>any</u> study in which the same individuals are observed at repeated intervals and in which nonresponse occurs. It can occur for example in the so-called complete follow-up surveys so frequently used to examine socioeconomic and medical characteristics. It is true that such studies give a detailed account of the history of the selected individuals, and that changes in the observation from T_1 to T_2 are certainly due to real changes in these individuals or at least to the response given by them. As we shall show, however, it is not necessarily appropriate to estimate changes in the population from the available observations on identical persons.

Also in this paper we undertake a comparison of estimates based on identical persons, on all persons, and on those persons interviewed only at one observation period. Since these comparisons appear to be awkward algebraically some numerical results are included.

Finally a simple "adjusted" estimation scheme is described. It is not recommended unreservedly however because it can exhibit poor behavior. The authors hope to report on additional work on the estimation problem in a future paper.

Since it was examination of the problem of systematic bias in unemployment statistics that led to the work described in this paper, we have chosen to describe the analysis in terms of employment and unemployment. However it is important to notice that none of the data in this paper are real and that the relationship of these models to the real problem has yet to be examined.

II. THE STATISTICAL MODEL

We have chosen to use a simple model for illustration. The increased algebra required for more general cases would detract from the presentation sufficiently that it seems better to avoid it. Consequently a two catagory model, employed and unemployed, is considered.

Suppose that a sample design results in the selection of a certain geographical area and that the employment status of all persons living in that area is to be determined both at time T_1 and again at a later time, T_2. For simplicity, we neglect the effect of population mobility. Many sample designs will specify a subsampling of persons in the selected area; the remarks we shall make apply to these cases also but for simplicity they are not discussed explicitly. Similarly, the higher structure of the sample design is also ignored with no loss in generality. Given these assumptions, all persons in the area to be sampled can be classified in the manner described in Table 2.1.

		Employment Status at Time T_2	
		Unemployed	Employed
Employment Status at Time T_1	Unemployed	N_{uu}	N_{ue}
	Employed	N_{eu}	N_{ee}

Table 2.1: Numbers of Persons Employed and Unemployed.

Using the notation of Table 2.1, the true unemployed/employed ratio at T_1 is given by

$$R_1 = \frac{N_{uu}+N_{ue}}{N_{eu}+N_{ee}} , \qquad (2.1)$$

and at T_2 by

$$R_2 = \frac{N_{uu}+N_{eu}}{N_{ue}+N_{ee}} . \qquad (2.2)$$

Algebraically, it can be seen that $R_1 = R_2$ if and only if $N_{ue} = N_{eu}$, that is, when the number of persons who have found employment in the period from T_1 to T_2 is equal to the number of persons who have lost it. Intuitively, this is an obvious condition for the unemployment ratio to remain constant.

When the actual sampling of the selected area begins, the classification of Table 2.1 will not be adequate, since not all persons will be interviewed. Consequently, each individual can be classified as employed, unemployed, or not interviewed at each of T_1 and T_2, as in Table 2.2. After the second round of interviews, the nine frequencies in this table will be known. We are assuming here that those persons who are not interviewed at either T_1 or T_2 can still be counted, so that the frequency F_{oo} is known. In some kinds of survey F_{oo} may remain unknown, or a rough estimate of it may become available. This has no effect on the subsequent discussion in this paper, which is directed towards estimation of the unemployed/employed ratios using the other eight frequencies, and which largely ignores sampling variations in the observed frequencies.

		Status at Time T_2		
		Unemployed	Employed	Not Interviewed
Status at Time T_1	Unemployed	F_{uu}	F_{ue}	F_{uo}
	Employed	F_{eu}	F_{ee}	F_{eo}
	Not Interviewed	F_{ou}	F_{oe}	F_{oo}

Table 2.2: Observed Numbers of Persons in Various Categories

To construct an elementary model, let

P_u = the probability that an unemployed person actually appears in the sample at T_1;

P_e = the probability that an unemployed person actually appears in the sample at T_1;

P_{uu} = the probability that an individual is interviewed at T_2 given that he was interviewed at T_1 and was unemployed at both T_1 and T_2;

P_{ue} = the probability that an individual is interviewed at T_2 given that he was interviewed at T_1 and was unemployed at T_1 and employed at T_2;

P_{eu} = the probability that an individual is interviewed at T_2 given that he was interviewed at T_1 and was employed at T_1 and unemployed at T_2;

P_{ee} = the probability that an individual is interviewed at T_2 given that he was interviewed at T_1 and was employed at both T_1 and T_2.

Finally, let Q_{uu}, Q_{ue}, Q_{eu} and Q_{ee} represent probabilities similar to P_{uu}, P_{ue}, P_{eu} and P_{ee} except that the Q's are conditional to the individual not being interviewed at T_1. For example Q_{uu} = the probability that an individual is interviewed at T_2 given that he was not interviewed at T_1 and was employed at both T_1 and T_2.

Ideally each of those probabilities would equal unity because all persons in the selected area are theoretically to be included in the sample. However non-response problems will virtually always ensure that these probabilities are not unity. Consequently it is interesting to construct a table of expected sample numbers based on the given three-by-three classification. These expectations are displayed in Table 2.3.

III. THE STUDY OF IDENTICAL PERSONS

The array of observed sample numbers Table 2.2, can be used to construct an estimator based only on individuals who

		STATUS AT TIME T_2		
		UNEMPLOYED	EMPLOYED	NOT INTERVIEWED
STATUS AT TIME T_1	UNEMPLOYED	$N_{uu}P_uP_{uu}$	$N_{ue}P_uP_{ue}$	$N_{uu}P_u(1-P_{uu}) + N_{ue}P_u(1-P_{ue})$
	EMPLOYED	$N_{eu}P_eP_{eu}$	$N_{ee}P_eP_{ee}$	$N_{eu}P_e(1-P_{eu}) + N_{ee}P_e(1-P_{ee})$
	NOT INTERVIEWED	$N_{uu}(1-P_u)Q_{uu}$ $+N_{eu}(1-P_e)Q_{eu}$	$N_{ue}(1-P_u)Q_{ue}$ $+N_{ee}(1-P_e)Q_{ee}$	$N_{uu}(1-P_u)(1-Q_{uu}) + N_{ue}(1-P_u)(1-Q_{ue})$ $+N_{eu}(1-P_e)(1-Q_{eu}) + N_{ee}(1-P_e)(1-Q_{ee})$

Table 2.3: Expected Sample Numbers

are observed at both T_1 and T_2. Using the table as a guide, it can be seen that the number of persons who are unemployed at T_1 and who are interviewed again at T_2 (i.e., in either the employed or unemployed category at T_2) is given by $F_{uu} + F_{ue}$. Similarly, the number of persons employed at T_1 and interviewed at T_2 is given by $F_{eu} + F_{ee}$. Consequently, the unemployment ratio at T_1, based only on those individuals who appear both at T_1 and T_2 is given by,

$$\hat{R}_1 = \frac{F_{uu}+F_{ue}}{F_{eu}+F_{ee}} . \qquad (3.1)$$

Similarly, the unemployment ratio at time T_2 for this same group of identical individuals is given by,

$$\hat{R}_2 = \frac{F_{uu}+F_{eu}}{F_{ue}+F_{ee}} . \qquad (3.2)$$

Furthermore, it can be seen that the difference $\hat{R}_1 - \hat{R}_2$ is given by,

$$\hat{R}_1 - \hat{R}_2 = (F_{ue}-F_{eu})(\text{positive factor}), \qquad (3.3)$$

so that $\hat{R}_1 = \hat{R}_2$ if and only if

$$F_{eu} = F_{eu}. \qquad (3.4)$$

In the usual survey situation, these frequencies will be large, so that under the model of Section II and neglecting sampling variability, (3.4) becomes

$$N_{ue}P_uP_{ue} = N_{eu}P_eP_{eu}. \qquad (3.4a)$$

In the case in which there is no overall change in employment, (i.e., $N_{ue} = N_{eu}$), this equation can be written as,

$$\frac{P_u}{P_e} = \frac{P_{eu}}{P_{ue}} . \qquad (3.5)$$

The result of Eq. (3.5) means that there will be a change in the observed unemployment ratio unless the ratio of the probability of interviewing an unemployed person to the probability of interviewing an employed person at, time T_1 is the same as the ratio of the corresponding probabilities at time T_2 for persons who were observed at T_1 and whose employment status has changed between T_1 to T_2. This change in the observed unemployment ratio will occur even though there is no change in the true unemployment ratio.

To explain this in a simpler case, consider the model in which the probabilities at T_2 are independent of the status at T_1 and of whether the individual is observed at T_1, as follows: $P_{uu} = P_{eu} = P_{2u}$, $P_{ue} = P_{ee} = P_{2e}$. Also we now write $P_u = P_{1u}$ and $P_e = P_{1e}$. Then Eq. (3.4a) can be written as,

$$N_{ue}P_{1u}P_{2e} = N_{eu}P_{1e}P_{2u}, \qquad (3.6)$$

and Eq. (3.5) as,

$$\frac{P_{1u}}{P_{1e}} = \frac{P_{2u}}{P_{2e}} . \qquad (3.7)$$

The interpretation of Eq. (3.7) is that the ratio of the probability of interviewing an unemployed person to the probability of interviewing an employed person must be the same at T_1 and T_2, otherwise there will be a change in the expected unemployment ratio from T_1 to T_2. This is true even though there has been no change in the true unemployment ratio and the estimates are based on identical individuals. Is Eq. (3.7) likely to hold in practice? It has been reported by Williams [10] that some practitioners feel that $P_{1u} > P_{1e}$. Consequently, if a survey manager subsequently reduces the nonresponse so effectively that P_{2u} and P_{2e} are approximately unity then the difference $\hat{R}_1 - \hat{R}_2$ must be positive and a systemmatic change in the estimator based on the identical persons will be observed. A similar result was also found by Williams [10] for estimators which are based on the complete sample at each of T_1 and T_2.

IV. A COMPARISON OF THE ESTIMATES BASED ON IDENTICAL, UNMATCHED, AND ALL PERSONS

In the previous section an analysis was made of the estimates based only on those persons appearing in the sample both at T_1 and T_2. Estimates can also be obtained which are based on those individuals who appear only once, either at T_1 and T_2. So altogether three pairs of estimates of R_1 and R_2 might be in hand; one based on identical persons, one on "single" persons and one on all persons. Consequently, it is interesting to ask whether all three of these estimates must necessarily exhibit a systematic change? It is possible for one estimate of the difference, $R_1 - R_2$, to behave differently from the others? And what are the magnitudes of the possible biases in the various estimates of $R_1 - R_2$?

The estimates based on identical individuals have already been described in

Section III. They were formed by simply picking out the appropriate sample numbers from Table 2.2. The estimate based on "single" individuals and the estimate based on all available persons can easily be formed in the same way. For example, the estimates based on the total number of persons, and their values under our model in the simple case of independence, are given by

$$\hat{R}_{1t} = \frac{F_{uu}+F_{ue}+F_{uo}}{F_{eu}+F_{ee}+F_{eo}} \approx \left(\frac{N_{uu}+N_{ue}}{N_{eu}+N_{ee}}\right) \frac{P_{1u}}{P_{1e}} , \tag{4.1}$$

and

$$\hat{R}_{2t} = \frac{F_{uu}+F_{eu}+F_{ou}}{F_{ue}+F_{ee}+F_{oe}} \approx \left(\frac{N_{uu}+N_{eu}}{N_{eu}+N_{ee}}\right) \frac{P_{2u}}{P_{2e}} . \tag{4.2}$$

The estimates based on the individuals who appear only on a single occasion are, (in the simple case of independence)

$$\hat{R}_{1s} = \frac{F_{ou}}{F_{eo}} \approx \frac{N_{uu}(1-P_{2u}) + N_{eu}(1-P_{2e})}{N_{eu}(1-P_{2u}) + N_{ee}(1-P_{2e})} \left(\frac{P_{1u}}{P_{1e}}\right) \tag{4.3}$$

$$\hat{R}_{2s} = \frac{F_{ou}}{F_{oe}} \approx \frac{N_{uu}(1-P_{1u}) + N_{eu}(1-P_{1e})}{N_{ue}(1-P_{1u}) + N_{ee}(1-P_{1e})} \left(\frac{P_{2u}}{P_{2e}}\right) \tag{4.4}$$

Unfortunately, the differences $\hat{R}_1 - \hat{R}_2$, $\hat{R}_{1t} - \hat{R}_{2t}$, and $\hat{R}_{1s} - \hat{R}_{2s}$ do not appear to have algebraic forms which facilitate easy comparisons. Consequently, a numerical study was undertaken. A large number of cases were computed; six are presented in the appendix. Each of the examples given there is arrayed in the same way, with the relevant population parameters followed by two tables. In the first table the expected sample numbers corresponding to Table 2.3 are laid out. Below this table, $E(n_1)$ and $E(n_2)$, the expected responses at T_1 and T_2 are listed. Table 2 contains the estimates of R_1, R_2, and $R_1 - R_2$ based on each of (i) identical persons, (ii) all persons (total) (iii) persons interviewed only at T_1 or T_2, (single) (iv) an adjusted estimator which will be described below.

Examination of the numerical results enables one to make certain comments:

1. Substantial biases appear with apparently innocous probability differences and with very low nonresponse rates. In example A, the response rate at T_1 is over 89 percent and at T_2, it is approaching 95 percent. These are response rates which are characteristic of some of the surveys run by the U.S. Bureau of the Census but are not matched consistently by any survey group. Nevertheless, the estimates based on all sampled persons suggest a change of about 14 percent in the unemployed/employed ratio. Keep in mind that $N_{eu} = N_{eu}$, so that there is no real change in unemployment, and also that sampling variability is being ignored. Even the estimate based on an identical set of individuals has a bias of about five percent. The estimates of R_1, R_2 and $R_1 - R_2$ based on the "singles" are so bad that it is disturbing to consider the possibility of their use.

Example B has been included to further illustrate the major effect that these probability differences can have on the estimates. The only difference between Example A and Example B is that the second stage response probabilities for unemployed persons, i.e., $P_{uu}, Q_{uu}, P_{eu}, Q_{eu}$ have dropped substantially. It is tempting to react to this case by taking the attitude that these probabilities are unrealistically low. But are they? And how would one know it, because the response rate at T_2 has dropped only one and one-half percent, which is very little different from Example A and in general is still a very high response rate. The important point however is that with almost no warning the biases in R_1, R_2 and $R_1 - R_2$ have gone from bad, in Example A to disastrous in Example B. The "total" change estimate now has a bias of over 55 percent! The estimates of R_1 and R_2 based on the identicals are clearly very bad, but in addition the "identical" estimate of $R_1 - R_2$ is now about 14 percent, or three times as big as it was in Example A.

2. Examples A and B have the same amount of shifting from one category to the other, i.e., $N_{ue} = N_{eu} = 100$ in both cases. This shifting does affect the biases. For example, suppose the probabilities of case A are used with the populations given below:

	N_{uu}	N_{ue}	N_{eu}	N_{ee}
i	400	00	00	9600
ii	300	100	100	9500
iii	200	200	200	9400
iv	100	300	300	9300
v	0	400	400	9200

These five examples are constructed so that the true unemployment ratio remains constant at 4.17%. It will be found that the biases in both the "total" and "identical" estimates get worse as the shifting around increases. On the other hand, the singles estimates get better. The latter are still not satisfactory but at least do not exhibit the extremely bad behavior of Examples A and B.

3. Calculation of case v in item 2 above reveals the distressing fact that the three estimates of the change $R_1 - R_2$ from identicals, total, and singles do not even necessarily have the same sign.

4. The Q probabilities do not seem to play a major role in the biases. These probabilities cannot affect the "identical" estimates at all and have only a small effect on the bias in the "singles" estimate. The estimates based on total persons are affected even more slightly because these are a combination of the single and the matched.

5. An examination of the tables of expected numbers reveals characteristics which are likely to be quite perplexing. For example the number of interviewed, employed, persons (at T_2), who were not interviewed at T_1 may be two or three times the number of persons who were employed at T_1 and not interviewed at T_2.

Specifically, in case A, the number of persons interviewed at T_2 and found to be employed, but who were not interviewed at T_1 is 999. This means that 999 employed persons who were not interviewed at T_1 were found in the sample area at T_2. On the other hand only 437 employed persons were interviewed at T_1 and lost to the sample at T_2. Thus it appears that one is "finding" far more employed persons at T_2 than were lost at T_1. This result might be interpreted as showing mobility of the population, so that unemployed persons move and show up somewhere else as an employed persons. Example A shows that this mobility feature may well be only a reflection of the response probability structure. This is not to say that the real population does not have such a characteristic, this example shows only that it will have to be demonstrated with other evidence.

6. The examples presented in the statistical appendix have what must be considered high expected response rates. If the probabilities are dropped so that the response rates are in the vicinity of 60 to 70 percent, the biases become far worse that those shown in cases A and B. Unfortunately, in practice, more surveys seem to operate in this range than with response rates of 90 percent.

7. Example C is constructed so that there is an even split of the population between the two categories. Again $N_{ue} = N_{eu}$ so that there is no change in the ratio in the two categories from T_1 to T_2. Since unemployment rates of this magnitude occur very seldom, it may be more helpful to discuss this example in terms of persons who say that they are going to vote for candidate U or for candidate E. The probabilities are numerically the same as those used in Example A. Now however they are interpreted as indicating that the voters for one candidate are more easily interviewed that those of the other. Both at T_1 and T_2 there is a response rate of approximately 91 percent which again can be considered to be high. However, the three estimates of $R_1 - R_2$, "identicals", "total" and "singles" give the appearance of a major swing to candidate E. Just how big a swing is determined by which of the three estimates is being considered. Unfortunately, in reality there is no swing to either of the candidates.

8. In practice the eight numbers (apart from the marginal totals) given in Table 1 of each of the examples are available for estimation purposes. Unfortunately our model has fourteen parameters, six P's, four Q's and four N's. In this situation, there are a number of simplifying assumptions that can be made to reduce the fourteen parameters to eight. One has been used in this paper. It consists of assuming that the four Q's are equal to the corresponding P's. In addition the assumptions $P_{uu} = P_{eu} = P_{2u}$ and $P_{ee} = P_{ue} = P_{2e}$ reduces the number of parameters to eight, specifically, two first stage P's, two second stage P's and four N's. Now the expressions for the expected cell numbers can be equated to the realized numbers and the resultant equations solved for the unknown eight parameters. This crude "adjusted" estimation scheme was used in all the examples shown in the appendix. While it behaves better than the other estimators in all of the present cases, and indeed did so in many of the computed examples which have not been included in the paper, it does not always behave so well. The estimation possibilities for these incomplete response cases required further investigation and the authors hope to report on this in a future paper.

V. COMPLETE FOLLOWUP STUDIES

Many sociological and medical surveys involve the repeated interviewing of all of the selected individuals. For example, a medical survey may involve the random selection of a group of people who are then all given periodic medical examinations. New persons may or may not be included in the survey at later stages, but to point out the potential difficulties in a simple way, suppose that no new individuals enter the study and that there is no loss from the original sample. Thus we put $P_{uu} = P_{ue} = P_{eu} = P_{ee} = 1$ and $Q_{uu} = Q_{ue} = P_{eu} = Q_{ee} = 0$ in the earlier models. While this model does not represent realistically the characteristics of sampling for unemployment statistics, for consistency we shall continue to refer to the classifications in that way. Given these assumptions the expectations are as given in Table 5.1.

		Status at Time T_2		
		Unemployed	Employed	Not Interviewed
Status at Time T_1	Unemployed	$N_{uu}P_u$	$N_{ue}P_u$	0
	Employed	$N_{eu}P_e$	$N_{ee}P_e$	0
	Not Interviewed	0	0	0

Table 5.1: Expected Sample Numbers, Complete Followups.

Consequently, at T_1,

$$\hat{R}_1 \approx \left(\frac{N_{uu}+N_{ue}}{N_{eu}+N_{ee}}\right)\frac{P_u}{P_e}, \qquad (5.1)$$

and at T_2,

$$\hat{R}_2 \approx \frac{N_{uu}P_u+N_{eu}P_e}{N_{ue}P_u+N_{ee}P_e}, \qquad (5.2)$$

Consequently,

$$\hat{R}_1 - \hat{R}_2 = (N_{eu}P_u - N_{eu}P_e)(\text{positive factor}), \qquad (5.3)$$

so that (neglecting sampling variability), $\hat{R}_1 = \hat{R}_2$ if and only if

$$N_{eu}/N_{ue} = P_u/P_e. \qquad (5.4)$$

From this result it can be seen that even when there is no overall change in "unemployment" from T_1 to T_2 (so that $N_{eu} = N_{ue}$), there will be change in the expectation of the estimator unless $P_u = P_e$. This rather disturbing result says that unless the probabilities of response at T_1 are known and are dealt with appropriately then the estimator may change from T_1 to T_2 even though the study involves an identical set of individuals at T_1 and T_2. There is no difficulty extending these results to followup studies with different and more complex sampling schemes.

Examples D, E and F have been included in the appendix to illustrate what may happen in the complete follow up case. Example D assumes that there is no response loss at T_2 and no new persons brought into the survey. The expected response is 89 percent. Even in this ideal case a one percent bias has crept in. In Example E, the second stage P probabilities have been lowered, so that some persons are lost from the survey but no new ones are added, i.e., the Q's remain equal to zero. The response rates are still high: 89 percent at T_1 and over 84 percent at T_2, nevertheless the biases in all of the estimates except the adjusted one have become unacceptably large. The "identical" change bias is about five percent and the "total" change bias is about ten percent. Since there are no persons appearing only at T_2, the singles estimate does not apply.

In Example F, P_{uu} and P_{eu} have been lowered substantially, to 0.50. It is important to notice that in practice there would be extreme difficulty in detecting whether one was more nearly in case E or F because the response rates are very nearly the same. The biases however have changed considerably. In case E, the biases were simply very bad, in case F they are much worse.

It is hoped that these simple examples will help to show the magnitude of the potential danger in panel surveys of this kind.

VI. SUMMARY AND GENERALIZATIONS

The important point made in this paper is that systematic changes will occur in estimates from one stage of a survey to the next if the probabilities of nonresponse are not the same for all of the population categories. Williams [10] showed that this was possible for the overall sample and in this paper the same systematic biases have been shown to be possible even if one computes estimates based on an identical set of individuals.

This paper also has included a discussion of complete followup surveys and a simple "adjusted" estimation scheme. Some numerical examples have been included which show the extreme biases that can easily be encountered in the estimates of change from one observation period of the next. This is true whether the estimate of change is based on identical persons, on all persons or on those persons appearing only for a single interview.

Distortions caused by the incomplete responses can mislead a researcher. For example, it is pointed out that a "mobility" characteristic could very well be a result of these nonresponse problems.

The model presented in this paper is a very simple one. It can be generalized easily to more categories and to additional reinterviews. With some increase in complexity the case of a continuous variate y_i can also be covered. We have chosen not to attempt extensive generalization in this paper. The list of cases that might be worth spelling out in detail is as long as the number of different schemes used in practice. It seems important simply to point out that these awkward biases can and probably do occur in practice.

REFERENCES

1. Hansen, M. H., Hurwitz, W. N. and Madow, W. G., Sample Survey Methods and Theory, John Wiley and Sons, New York, Vols. I and II, 1953.

2. Cochran, W. G., Sampling Techniques, Second Edition, John Wiley and Sons, New York, 1963.

3. Patterson, H. D., Sampling on successive occasions with partial replacement of units, Jour. Roy. Stat. Soc., B, 12, 241-255, 1950.

4. Eckler, A. R., Rotation Sampling. Ann. Math., Stat., 26, 664-458, [1955].

5. Rao, J. N. K., and Graham, Jack E., Rotation designs for sampling on repeated occasions. Jour. Amer. Stat. Assoc. 59, 495-509, 1964.

6. Kish, Leslie, Survey Sampling, John Wiley and Sons, New York, 1965.

7. U.S. Bureau of the Census, The Current Population Survey - A Report on Methodology. Technical Paper No. 7, U.S. Government Printing Office, Washington, D.C., 1963.

8. Waksberg, Joseph, and Pearl, Robert B., The effects of repeated interviews in the current population survey. Paper presented at the 47th National Conference of the American Marketing Association, Dallas, Texas, 1964.

9. Measuring Employment and Unemployment, Report of the President's Committee to Appraise Employment and Unemployment Statistics. The White House, 1962.

10. Williams, W. H., The systematic bias effects of incomplete responses in rotation samples. To appear in the Public Opinion Quarterly.

11. Durbin, J., and Stuart, A., Differences in response rates of experienced and inexperienced interviewers. Jour. Roy. Stat. Soc., 114, 164-206, 1951.

APPENDIX

EXAMPLE A

First Stage Response Probabilities

$P_u = 0.93$ $P_e = 0.89$

Second Stage Response Probabilities

$P_{uu} = 0.87$ $P_{eu} = 0.96$ $P_{eu} = 0.84$ $P_{ee} = 0.95$

$Q_{uu} = 0.87$ $Q_{ue} = 0.96$ $Q_{eu} = 0.84$ $Q_{ee} = 0.95$

True Population Figures

$N_{uu} = 300$ $N_{eu} = 100$ $N_{eu} = 100$ $N_{ee} = 9500$

TABLE 1: EXPECTED SAMPLE NUMBERS

	U2	E2	N2	T2
U1	243	89	40	372
E1	75	8032	437	8544
N1	28	999		
T1	345	9121		

$E(n_1) = 8916$
$E(n_2) = 9466$

TABLE 2: ESTIMATES OF UNEMPLOYMENT RATIOS (PERCENT)

	TRUE	IDENTICALS	TOTAL	SINGLES	ADJUSTED
R1	4.17	4.10	4.35	9.15	4.17
R2	4.17	3.91	3.78	2.75	4.11
R1 - R2	0.00	0.19	0.57	6.40	0.05

EXAMPLE B

First Stage Response Probabilities

$P_u = 0.93$ $P_e = 0.89$

Second Stage Response Probabilities

$P_{uu} = 0.50$ $P_{ue} = 0.96$ $P_{eu} = 0.50$ $P_{ee} = 0.95$

$Q_{uu} = 0.40$ $Q_{ue} = 0.96$ $Q_{eu} = 0.40$ $Q_{ee} = 0.95$

True Population Figures

$N_{uu} = 300$ $N_{ue} = 100$ $N_{eu} = 100$ $N_{ee} = 9500$

TABLE 1: EXPECTED SAMPLE NUMBERS

	U2	E2	N2	T2
U1	139	89	143	372
E1	44	8032	467	8544
N1	13	999		
T1	197	9121		

$E(n_1) = 8916$

$E(n_2) = 9318$

TABLE 2: ESTIMATES OF UNEMPLOYMENT RATIOS (PERCENT)

	TRUE	IDENTICALS	TOTAL	SINGLES	ADJUSTED
R1	4.17	2.83	4.35	30.65	4.08
R2	4.17	2.27	2.16	1.28	4.08
R1 - R2	0.00	0.57	2.20	29.37	-0.01

EXAMPLE C

First Stage Response Probabilities

$P_u = 0.93$ $P_e = 0.89$

Second Stage Response Probabilities

$P_{uu} = 0.87$ $P_{ue} = 0.96$ $P_{eu} = 0.84$ $P_{ee} = 0.95$

$Q_{uu} = 0.87$ $Q_{ue} = 0.96$ $Q_{eu} = 0.84$ $Q_{ee} = 0.95$

True Population Figures

$N_{uu} = 4000$ $N_{ue} = 1000$ $N_{eu} = 1000$ $N_{ee} = 4000$

TABLE 1: EXPECTED SAMPLE NUMBERS

	U2	E2	N2	T2
U1	3236	893	521	4650
E1	748	3382	320	4450
N1	336	485		
T1	4320	4760		

$E(n_1) = 9100$

$E(n_2) = 9080$

TABLE 2: ESTIMATES OF UNEMPLOYMENT RATIOS (PERCENT)

	TRUE	IDENTICALS	TOTAL	SINGLES	ADJUSTED
R1	100.00	99.99	104.49	162.55	100.00
R2	100.00	93.20	90.76	69.25	97.63
R1 - R2	00.00	6.79	13.74	93.30	2.37

EXAMPLE D

First Stage Response Probabilities

$P_u = 0.93$ $P_e = 0.89$

Second Stage Response Probabilities

$P_{uu} = 1.00$ $P_{ue} = 1.00$ $P_{eu} = 1.00$ $P_{ee} = 1.00$

$Q_{uu} = 0.00$ $Q_{ue} = 0.00$ $Q_{eu} = 0.00$ $Q_{ee} = 0.00$

True Population Figures

$N_{uu} = 300 \quad N_{ue} = 100 \quad N_{eu} = 100 \quad N_{ee} = 9500$

TABLE 1: EXPECTED SAMPLE NUMBERS

	U2	E2	N2	T2
U1	279	93	0	372
E1	89	8455	0	8544
N1	0	0		
T1	368	8548		

$E(n_1) = 8916$

$E(n_2) = 8916$

TABLE 2: ESTIMATES OF UNEMPLOYMENT RATIOS (PERCENT)

	TRUE	IDENTICALS	TOTAL	SINGLES	ADJUSTED
R1	4.17	4.35	4.35	0.00	4.35
R2	4.17	4.31	4.31		4.31
R1 - R2	0.00	0.05	0.04		0.05

<u>EXAMPLE E</u>

First Stage Response Probabilities

$P_u = 0.93 \qquad P_e = 0.89$

Second Stage Response Probabilities

$P_{uu} = 0.87 \; P_{ue} = 0.96 \; P_{eu} = 0.84 \; P_{ee} = 0.95$

$Q_{uu} = 0.00 \; Q_{ue} = 0.00 \; Q_{eu} = 0.00 \; Q_{ee} = 0.00$

True Population Figures

$N_{uu} = 300 \quad N_{ue} = 100 \quad N_{eu} = 100 \quad N_{ee} = 9500$

TABLE 1: EXPECTED SAMPLE NUMBERS

	U2	E2	N2	T2
U1	243	89	40	372
E1	75	8032	437	8544
N1	0	0		
T1	317	8122		

$E(n_1) = 8916$

$E(n_2) = 8439$

TABLE 2: ESTIMATES OF UNEMPLOYMENT RATIOS (PERCENT)

	TRUE	IDENTICALS	TOTAL	SINGLES	ADJUSTED
R1	4.17	4.10	4.35	9.15	4.35
R2	4.17	3.91	3.91		4.25
R1 - R2	0.00	0.19	0.44		0.10

<u>EXAMPLE F</u>

First Stage Response Probabilities

$P_u = 0.93 \qquad P_e = 0.89$

Second Stage Response Probabilities

$P_{uu} = 0.50 \; P_{ue} = 0.96 \; P_{eu} = 0.50 \; P_{ee} = 0.95$

$Q_{uu} = 0.00 \; Q_{ue} = 0.00 \; Q_{eu} = 0.00 \; Q_{ee} = 0.00$

True Population Figures

$N_{uu} = 300 \; N_{ue} = 100 \; N_{eu} = 100 \; N_{ee} = 9500$

TABLE 1: EXPECTED SAMPLE NUMBERS

	U2	E2	N2	T2
U1	139	89	143	372
E1	44	8032	467	8544
N1	0	0		
T1	184	8122		

$$E(n_1) = 8916$$

$$E(n_2) = 8306$$

TABLE 2: ESTIMATES OF UNEMPLOYMENT RATIOS (PERCENT)

	TRUE	IDENTICALS	TOTAL	SINGLES	ADJUSTED
R1	4.17	2.83	4.35	30.65	4.35
R2	4.17	2.27	2.27		4.29
R1 - R2	0.00	0.57	2.09		0.06

Discussion

Walt R. Simmons, National Center for Health Statistics

1. *Introduction*

My discussion will focus mainly on the paper by Dr. Greenberg, et. al. At the close, I shall make two brief comments on the Mallows and Williams paper.

I must warn anyone in the audience who is not already aware of it that on the subject of randomized response I am a prejudiced witness. Ever since I first heard of the original Warner proposal, I have been intrigued by the notion. I have explored a number of theoretical variations of the original model, have talked with quite a few people—including the authors of the present paper—written about advantages and weaknesses of the method, and have had a role in several of the field trials that have been undertaken. Out of this experience, there are three principal conclusions which I should like to pass on:

First: "Randomized Response" is not a single technique, but rather under its umbrella there can be assembled an *extensive family of procedures,* characterized by the fact that they can provide essentially unbiased estimators even though no person other than the respondent himself knows or possibly *can* know how the individual respondent replied to a specified question.

Second: At least some of these procedures are operationally feasible, and are efficient devices for securing intelligence on sensitive issues.

Third: The efficiency of one randomized response model can vary greatly from that of another. Nor is it always initially obvious which of two procedures is better, as was discovered when we first explored the weaker-appearing unrelated question version, only to find it could have smaller variance than certain competitors. The total process, being a function of at least several parameters, is sufficiently complex that it merits and repays the expenditure of considerable effort to locate optimum formulations.

My bias clearly, then, favors randomized response methods, and I am grateful to the authors for their work in advancing the technique. Please remember that overall judgment as I use most of my time to call attention to difficulties, weaknesses, or problems associated with the scheme described in today's paper, or to the method more broadly.

2. *New Contributions*

The principal contributions of the present paper are two in number.

On the theoretical side the approach has been extended for the first time to quantitative variables—e.g., amount of income of head of household—whereas previous work has been restricted to binomial, yes-no, true-false situations. This is obviously a significant extension. The theoretical work follows largely the pattern of earlier treatment of the binomial. The audience may feel that the presentation in this paper lacks some of the satisfying features of the earlier work—especially algorithms for optimum determination of design parameters: choice of the y-statistic given a likely range for the A-statistic; apportionment of n into n_1 and n_2; and choice of P_1 and P_2. Apparently the actual design work did follow the guidelines of the earlier binomial analysis.

The authors recognize in the empirical data evidence counter to the assumption that the A-statistic and the y-statistic have the same probability density function. This is not important in calculation of the key estimator, since that computation is distribution-free. But it can have an impact on the climate of the interview, and on the optimum determination of design parameters. Sensitivity of design analysis to operational deviations from the assumption of equivalent distributions should be investigated further. Remember that we rarely know precisely the density function of the A-variable, and may very well not know that of the y-variable.

The second contribution of the paper is of course the presentation of data from field trials. This is a very important part of any theoretically contrived procedure such as randomized response. One must accumulate experience to determine if the method is feasible. The authors report that the operation was a feasible one, and argue that the results are plausible.

3. *Respondent Cooperation*

Use of randomized response technique is designed to minimize non-response and evasive or incorrect response. Consequently, the character of the approach to the respondent is important. Presentation time today was limited, but most serious audiences would like to know precisely how the game was explained to the respondent, and how the interviewer went about convincing the respondent to tell the truth.

One very important aspect of this matter is that of assuring that the respondent truly understands the procedure and the specific questions. The procedure *is* unusual, must in some degree cause diversion of attention of the respondent while he tries to discover what the trick may be. One wishes to forestall that line of thinking if possible.

Question construction can be critical. In an earlier survey, one question read, "There was a baby born in this household after January 1, 1965, to an unmarried woman who was living here." This is a bit more involved question, than one would prefer. There is a suspicion that some respondents missed the *un* in "unmarried," and replied as though the question had asked about children of married women—a more normal idea perhaps.

In the first of the two studies reported in today's paper, the sensitive issue was abortions. The two questions were:

(1) How many abortions have you had during your lifetime?
(2) If a woman has to work full time to make a living, how many children do you think she should have?

Dr. Greenberg has said they "goofed" in not making it clear that the second question related only to married women. I guess that is so, but the unrelated question seems to me otherwise to be particularly well-chosen. Did the framers consciously select a statement which had propaganda value in order to encourage the respondent on the acceptability of abortion, and therefore to answer question 1 truthfully?

4. *Realized Values of P_i*

The unbiased estimator of the sensitive statistic is

$$\bar{A} = \frac{(1 - P_2)\,\bar{z}_1 + (1 - P_1)\,\bar{z}_2}{P_1 - P_2}$$

The $\bar{z}_i$ value is the observed mean of responses in the i^{th} sample, and P_i is the designed proportion of instances in which the sensitive question is asked in the i^{th} sample.

If we start with the values

$$\mu_A = 0.5$$
$$\mu_y = 0.9$$
$$P_1 = 0.7$$
$$P_2 = 0.3$$

following the paper, we get an expected sample estimate for μ_A of $\bar{A} = 0.5$ as we should hope. Suppose, however, that the *realized* values of P_i, secured from faulty construction or handling of the randomizing device, are $P_1' = 0.5$ and $P_2' = 0.1$. Then we should expect an estimate $\bar{A}' = 0.58$, a substantial error. The error from this source tends to be less when μ_A and μ_y are similar, and greater when they are dissimilar. The possible failure of the realized proportion of persons answering question 1 to match the designed probability P_1 represents one of the important inherent technical—or perhaps I should say technological—frailties of the system. Perhaps pilot tests of the randomizing procedure should be made until a non-biased process can be reasonably guaranteed. Perhaps some new technique should be invented which will assure that the realized and designed values P_i are closely similar.

5. *Empirical Evaluation*

Some sturdier empirical tests or evaluations would be useful. We shall agree readily with the authors that in some respects, including directions of observed differentials, results of both of the studies are plausible. Although, uninformed of research in the area and recognizing that a majority of respondents are below the mean income, I'm not sure I would concur with the authors in expecting that both white and nonwhite respondents would report that the average head of the household of the same size had higher income than their own household.

The procedure estimates $\bar{y}$ as well as $\bar{A}$, the sensitive measure. In certain situations it should be feasible to choose a y-statistic for which the mean is known. In this way one can secure an external check on the survey. In fact, there may exist decent external statistics on income of persons in the counties in which this survey was conducted. In a hasty review I discovered some IRS data for North Carolina which suggest that the report figures on income are at least in the right ball-park.

6. *Sampling Variances*

The paper states that the samples were stratified cluster designs, but that sampling variances were calculated on the assumption of simple random sampling of z-values. I would suspect that there was a more-than-trivial intracluster correlation on the income variables and perhaps on abortion. If so, I suspect the authors will agree that the variances shown in the report are understatements.

7. *The Future*

I hope other variants of the general technique will be developed, and other trials made of the procedures already described. These can take several different paths. I'd like to suggest one effort: Namely, a device which will adapt randomized response to a mail survey. One very important proviso goes with this. That is that it shall indeed continue to be true that it is *impossible* for the surveying agency to know whether or how the respondent has replied to a specified question.

8. *The Mallows and Williams Paper*

There are two remarks I should like to make on this report. Basically, the major part of the paper is a rather elaborate algebraic description of a model which demonstrates that biased measurements will result from a survey unless non-response rates for relevant subcategories are equivalent. We will all agree with this finding. It is why most of us attempt to impute for non-response within subclasses which are as nearly homogeneous as we can make them.

The Mallows and Williams paper offers one technique among a number of possible alternatives for adjustment of raw data. I am certainly among those who believe that we should do everything we can to minimize the impact of non-response. That task will be difficult unless the extent of non-response is kept small. The old adage applies: It's better to do analysis with data than without.

Discussion

Benjamin J. Tepping, Bureau of the Census

First, I shall comment briefly on the very interesting paper presented by Dr. Greenberg.

The assumption that the probability distribution associated with the sensitive question is the same as that associated with the non-sensitive question, except for the location parameter, is not an appealing one, for one may expect that in practice the distributions may be quite unlike each other. Moreover it is easy to see that the unbiased estimator can be obtained without any assumption regarding the forms of the distributions.

It follows that, while an investigation of the maximum likelihood estimator could yield some information about the relative efficiency of the unbiased estimator, the maximum likelihood estimator is not to be preferred in practice, since it would require assuming the functional form of the probability distribution. On the other hand, an interesting investigation from the practical point of view would be that of the optimum allocation of the design parameters, given that the unbiased estimator is to be used. Another interesting question concerns the relative magnitudes of the mean-square errors for (a) the direct interview approach for a single question, (b) the randomized response approach assuming that all response biases eliminated, and (c) the randomized response approach if one assumes that the bias is only reduced by specified amounts.

The valuable paper by Mallows and Williams is concerned with an extremely serious problem. As the authors have noted, this bias (which we have come to call the "rotation group bias") has been observed in the Current Population Survey conducted by the Bureau of the Census and in one or two other panel surveys. In fact, however, this bias probably exists in <u>every</u> survey, whether it is a survey based on a fixed panel of respondents or a survey that uses rotating subsamples. It probably exists even in a one-time survey, since clearly it can be regarded as the first interview of a panel survey. Of course, it is only in the latter case that the bias can be detected clearly, and then only if the several rotation groups for each round of the survey are separately tabulated. Thus it is important, for <u>all</u> surveys, to understand the causes of the bias, for such an understanding may lead to the adoption of procedures for reducing these biases.

Those who design or conduct surveys and those who make use of survey results will therefore welcome the discussion presented by Mallows and Williams, in which the bias can be shown to appear if there is a differential probability of being interviewed for persons designated for the sample. We cannot, however, regard it as established that the differential probability of being interviewed does <u>in fact</u> account for any substantial part of the bias that has been observed. The hypothesis advanced by Mallows and Williams is one of a dozen or more hypotheses or classes of hypotheses that have been advanced in attempts to explain the rotation group bias. I wish briefly to mention a few of the other hypotheses.

It may be that one contribution to the bias is the fact that attitude and the behavior of the interviewer varies as a function of the length of time that the sampled unit has been a member of the sample. It may be that the interviewer is required to obtain more detailed information in the course of the first interview, for example demographic information about each member of the household. It is quite possible that the interviewer may be quite meticulous in his questioning of a new unit, but become less careful after repeated interviews, perhaps by assuming the answers to questions that he does not in fact ask.

Interviewers are not assigned their work at random . and the interviewers that are assigned to units new to the sample may tend to be of a different type, perhaps because it may seem desirable to the supervisors to assign particularly skilled interviewers to new cases. This policy may contribute to a difference between units new to the sample and those that have previously been interviewed.

In the Current Population Survey, for example, certain probing questions are asked of persons whose initial responses during the interview would indicate that they are not in the labor force. These probes result in some of these persons being classified finally as unemployed. But the probes are used only for units of the first and the fifth months in the sample, thus possibly contributing to the bias observed.

The identity of the respondents actually reporting for the given household in the sample varies over the life of the panel, and the characteristics of the respondent and his relationship to the person about whom information is recorded affect the responses given. If, as is usual, the respondent is not chosen at random, this could give rise to a rotation group bias.

The bias may be the result of psychological conditioning by means of previous interviews, without any real changes in the characteristics that are to be reported.

It may be that some respondents will merely recall and repeat an earlier response. The hypothesis has been advanced that the probability of such behavior is related to the characteristics of the respondent. It can be shown that this would lead to a rotation group bias.

Finally, we cannot overlook the possibility that the characteristics of some persons may actually change as a result of being interviewed.

For example, a person originally not in the labor force may begin looking for work, or an unemployed person may leave the labor force, as a result of his labor force status having been discussed during previous interviews.

It is likely that more than one of the possibilities that have been mentioned (or some that have not been mentioned) are effective to some degree. A considerable amount of research has been done through the investigation of existing data and even by mounting special experiments to measure the effects implied by the hypotheses mentioned. A great deal more research is probably required to determine the most important causes of the rotation group bias, and to find ways of reducing or, hopefully, eliminating the bias.

V

LUNCHEON ADDRESS

Chairman, HENRY S. SHRYOCK, U. S. Bureau of the Census

FUTURES RESEARCH: A NEW SCIENTIFIC DISCIPLINE?

Frank P. Davidson, The Institute for the Future

It was perhaps no mere coincidence that one week before Armstrong and Aldrin landed on the moon, President Nixon announced the establishment, within The White House, of a National Goals Research Staff. The Apollo program illustrates the dramatic enlargement which our generation has witnessed in the scope of "the feasible." Faced with the consequent plethora of alternative futures, choices among them -- if they are to be solidly based on preferred values, presuppose careful and logical analysis of the implications and consequences of alternative policy decisions.

The Institute for the Future (IFF) was established a year ago to institutionalize systematic and comprehensive studies of the long-range future. IFF thus became America's first organization expressly designed to analyze the long-term implications of today's major policy decisions. As such, it assumed the task of forecasting technological, social and economic developments in order to identify workable and desirable departures from established trends. IFF does not presume to "predict the future" but only to identify and assess alternative futures in terms of their social costs and benefits, and so to contribute to the better definition of national goals and programs.

As a priority task, the staff of IFF is endeavoring to evaluate and refine methodologies of forecasting. But as early as November, 1967, when our initial prospectus was written, IFF accepted that "it is not enough to simply forecast, research and report... it is essential that as many key executives as possible -- in government, business, universities and foundations-- develop the competencies required to utilize the reports and services of forecasting centers." Therefore, we were delighted when President Nixon, in his statement of July 13, 1969, underlined the "urgent need to establish a more direct link between the increasingly sophisticated forecasting now being done and the decision-making process." And, the President added, "our need now is to seize on the future as the key dimension in our decisions, and to chart that future as consciously as we are accustomed to charting the past."

Statisticians readily appreciate the inadequacy of extrapolation from the past as a method of forecasting future developments, especially on a long-range basis. So much of our economy depends on the evolution of science and invention that responsible forecasting efforts must include data as to new technological developments and, equally, the reactions of ordinary people to technological change and its social consequences. We must, somehow, account for the baffling discontinuities of modern life!

But to develop responsible and - within limits - reliable forecasts requires judicious combinations of innovative methods and approaches. Among these is the Delphi technique described in Olaf Helmer's <u>Social Technology</u>.[1] Helmer and his colleagues found that, in the average meeting, experts tend to react to each other, to the detriment of concentration on and refinement of the exact questions at issue. Whereas, in Delphi studies, successive questionnaires are sent to panels of carefully selected specialists; after several rounds of expert judgments from the respondents, a consensus can often be derived as to the probabilities of future events and their impacts on society.

IFF has completed a number of Delphi studies devoted to anticipated developments in physical technology, bio-medicine and social change. Of course, this method differs fundamentally from statistical sampling, where it is important that pollsters have the same distribution in the sample as in the population as a whole, and the sample must duplicate the larger group in important respects so that the profile of the smaller group represents the profile of the larger. However, Delphi studies rely on expert judgment and seek consensus from those best able to make disciplinary contributions to a central question. The disciplines involved may, in a given study, overlap or be adjacent, but must be combined in some explicit way before there can be confidence that a complete answer is at hand. The Delphi procedure was developed precisely because scientific questions cannot be decided by majority vote, even among scientists! What is wanted is a logical process of debate so that mutual criticism and stimulation can refine issues and reduce ambiguities. Where, nonetheless, ambiguities persist, the feedback process is useful in crystallizing opposing points of view.

But an agreed list of probable future developments can, itself, be a source of confusion. There is a need for systematic analysis of the inter-actions of the forecasted events. One event may enhance or impede the emergence of others, and the effects of one on another may be slight or immense, as well as immediate or remote in point of time. Accordingly, T. J. Gordon has developed a "cross impact matrix" method of forecasting the interactions of future events.[2] This method shows great promise as a means of testing and refining the results of Delphi studies.

IFF has also developed simulation models so that planners and decision-makers can better visualize the implications of policy and investment decisions. For instance, under a grant from the Connecticut Research Commission, and using Delphi and cross-impact techniques, IFF designed a simulation game of the future economy of the State of Connecticut. Under a grant from Wesleyan University, we are undertaking preparatory studies for an urban simulation laboratory.

Work is going forward on the design of games simulating corporate decision-making, as well. And in-house programs are currently devoted to the advance of the theory of social indicators, analysis of the so-called quality of life and clearer definition and measurement of societal values. There are many statistics about society but often these are not in useful form for assessing the "health" of society. The movement for utilizing statistics as a basis for understandable social indicators is therefore one in which your participation is urgently needed, as a basis for better assessment of the nation's present position and its plans for the future.

We are also studying the development and use in forecasting of electric communications facilities to interconnect a world-wide network of scholars. This "D-Net" will be an on-line interactive group communications system which will make it possible for decision-makers to focus expert judgment on some of the more urgent problems of today. Our basic research, then, has two major components:

> development of forecasting methods and other tools for the analysis and synthesis of potential futures, and
>
> the application of such techniques to the problems of society.

We concede that studies of potential futures are, in and of themselves, an insufficient response to the greater problems of our age. A common characteristic of today's social problems is that their solution requires not only multidisciplinary intellectual in-puts but also inter-sectoral social action. The sectors of our society usually designated as "public" and "private" need improved methods of collaboration, if either sector is to accomplish its job effectively.

We must, therefore, accept the necessity for innovative relationships between "knowledge" and "authority." Numerous expert studies which are not applied - but could or should be - only add to the disparity between our potential for impressive social achievement and the persistent and embarrassing realities of urban blight, atmospheric pollution, road and airport congestion, the alienation of youth, and other well-publicized ills of our time. Even in a "knowledge society," knowledge is not automatically translated into effective decisions. There is an institutional gap, often the more insidious because it remains unrecognized. We need to pay increased attention to the institutional mechanisms whereby feasible and desirable plans can enter the mainstream of life. Harold Lasswell, of Yale, has recommended "decision seminars," - study groups whose members represent both expert knowledge and the authority to apply it in defined areas. For this idea, there are useful precedents: the Channel Tunnel Study Group, for example, combined governmental and private interests in a common program to establish a permanent land link between Britain and France. In our own country, COMSAT demonstrated the important services which can be rendered by a "mixed economy" corporation based on the latest state-of-the-art technology. But isolated achievements should not blind us to the fact that we lack a whole range of institutions capable of transforming social research into social achievement. The recently established New York Urban Development Corporation was one response to this problem of the "institutional gap."

How can we encourage a new union of intellectual and financial resources, a union that shuns outmoded disciplinary attitudes and sectoral dogmas and yet retains a genuine capability for concerted and useful action?

One approach, I suggest, is the establishment of study groups in which systems analysts and other experts can meet informally with representatives of the financial and governmental institutions which must make or approve major policy and investment decisions. This may be a fundamental step if American society is to achieve the quality of life which all of us now recognize as being within our grasp. One such study group recently resulted in the organization of the "Geo-Transport Foundation of New England, Inc." which will test the willingness of industry and government to invest cooperatively in a modern inter-urban transport system between New York and Boston. For a highly developed country such as ours, conditions on the New Haven Railroad constitute an anachronism. No effective center of decision or responsibility exists: the railroads alone cannot provide the required level of service; governmental interests are fragmented among numerous Federal, state and municipal authorities, and industry has not had a clear channel for the coordination of plans and programs that would provide an essential element of the future infra-structure of New England and New York. The new Geo-Transport Foundation, which groups support from public and private sectors, may serve as a model for "combined operations" in many fields where social and private interests alike require agreement on choice of system and choice of investment program.

A sober respect for statistical reality must, of course, be a constituent element of any scheme for institutional engineering. But the forecasting business has peculiar hazards: just as "systems analysis" offers an approach to decision-making under conditions of uncertainty, so "institutional engineering" must allow for flexibility and adjustment as conditions change and knowledge improves. The study group, in this sense, may be thought of as a device to avoid an over-rigid organizational form; it is a step preliminary to the formation of a public authority, a mixed economy corporation, a joint venture, or some other appropriate institution.

So conceived, study groups might well be established to assess the cooperative use and development of continental water resources; the financing on a more systematic basis of our educational infra-structure; the adaptation of information systems to the needs of education; the future of the Bering Straits; the long-range control of environmental pollution, and a host of other recalcitrant problems.

Just as PPBS represented an advance over traditional departmentalized concepts of government administration, so the inter-sectoral study group can be useful as a procedural device to marshall community resources on behalf of agreed national and international goals. Indeed, the inter-sectoral study group may be viewed as a step beyond the "think tank" -- as an embryo "action tank" capable, in time, of translating systems analysis into the systems and institutional arrangements of the future. Such a development, broadly conceived, could constitute a fundamental contribution to the program for identifying and achieving realizable national goals. In the resulting social context, both "forecasting" and "inventing" alternative futures could lose any residual tinge of "gamesmanship" and enter, at last, the valhalla of scientific respectability.

[1] Social Technology, Olaf Helmer, Basic Books, New York, 1966.

[2] "Initial Experiments with the Cross Impact Matrix Method of Forecasting," Futures, Vol. 1, No. 2, December 1968.

VI

COMPUTERS IN SOCIAL SCIENCE DATA PROCESSING AND ANALYSIS

Chairman, HUGH F. CLINE, Russell Sage Foundation

INTERACTIVE GRAPHIC CLUSTERING USING THE PROMENADE SYSTEM

David J. Hall, Geoffrey H. Ball, and Daniel E. Wolf
Stanford Research Institute

1. Introduction

The PROMENADE system is used to explore data using a number of manipulative, graphic, and numeric programming subsystems that are linked together under a single executive program. Interactive interface between subsystems is as important as interactive interface between computer and user.

Principal features of the system are convenience, flexibility, and on-line interaction. Facilities are available for keeping a record of the analysis, using the line printer, still and movie cameras, and a 35-mm microfilm recorder.

This approach to clustering stresses the use of multivariate plots that allow the user, for example, to "fly through hyperspace," observing his data. In addition, multivariate histograms, scatter plots, link-node plots, waveform plots, and others can be dynamically controlled from the console to provide a unique visual insight into data structure, which cannot be obtained from conventional off-line media.

Numeric algorithms that have been developed at SRI are also integrated into the system. One is for cluster analysis and the other for pattern recognition, and both can be interactively controlled from the console. This paper will not discuss the use of the pattern-recognition algorithm. The clustering method known as ISODATA has been available in the system for at least three years. However, as a research topic itself, apart from the existence of PROMENADE as a separate research topic, it is in a fluid state of continual change--hopefully, effecting improvements. The recent additions to the clustering system have been to allow it to be controlled directly from two graphic subsystems in addition to the usual mode of control through its own executive. These three modes of control are all interactive, with the two graphic modes being the most interactive and allowing control of extra details. When the title for this paper was selected, we had hoped to have, at the final writing, some extensive experience at controlling ISODATA from the graphics. However, there were delays in mapping our software system to a new machine configuration, and this paper will therefore be less detailed in this respect. It is also necessary to explain other related features of PROMENADE, before describing interactive control from the graphic subsystems.

Manipulative control of the data is afforded by another subsystem responsible for reading data, in almost any format, and storing it on the disk. This subsystem also provides for selection of a subset of variables or patterns, for performing elementary transformations, and for formation of derived variables, in a convenient way using the keyboard.

The software system structure, using disk overlays, allows the indefinite expansion of the system to eventually incorporate a wider range of perhaps more conventional subsystems and associated program interfaces.

It is predicted that statistical data analysts will prefer on-line interactive graphic analysis in the future. The reasons for this preference will be explained in the following sections.

The PROMENADE system is currently being used on four significant applications:

(1) Chemical--classification of spectrometer data

(2) Marketing--clustering of data on companies

(3) Census--price-index evaluation for housing construction data

(4) Cloud clustering of satellite weather photographs.

a. The Status of Computers and User Modes of Operation

Computers seem to be here to stay, and we have to learn to live with them and make the best use of this powerful tool. So the question is not a matter of choosing between employing computers or not employing them in our statistical data analysis work, but of finding the mode of operation with computers that is best suited to our applications, and the programs that can help us solve our own particular problems. As far as the user operating mode is concerned, there are some broad categories:

(1) Batch computing, characterized by the complete specification, before the computer run, of any control quantities required by the program. A user of a batch program is generally allowed no interaction during the running of the program and should be sufficiently confident of the nature of the program that he is running, probably having used it several times before. The batch user probably requires copious printout, only a small fraction of which is actually read. Most of this printout is obtained in case some additional information is needed.

(2) Interactive, or conversational computing, in which the user works at a teletype keyboard. The user can interrupt the process, or feed it new data while it is in progress, and usually operates in a timesharing computer system. This usually implies a low printout rate of approximately ten characters per second.

(3) Interactive computing, but with graphics, in a dedicated computer

system--i.e., not time-shared. This mode of operation generally implies a high dollar cost per hour for use of the system because while the user is on-line, he is using the whole computer system including card reader and punch, tape units, and line printer in a manner that does not allow sharing of this equipment. Our own facility, a CDC 3300, is now operating in this mode. It is planned to operate in a time-sharing mode early in 1970, but meanwhile, the user of the system must pay approximately $100 per hour.

(4) Interactive graphics with time-sharing, which is a fairly rare commodity but much to be desired. A good system will have sufficient graphical power and relatively rapid response as far as the human is concerned, and because the computer system is being shared, lower costs are incurred than for the dedicated system. Such systems frequently operate around a computer center, with the consoles in fairly close proximity.

(5) Remote graphics computing. A console providing graphic capability is connected to a distant computer, perhaps across the country. This is not generally available today; however, we must expect that such facilities will eventually be available, and plan our data-analysis strategies for the future with this type of network facility in mind. Possible a small computer at the same site as the display console would be an ideal arrangement.

b. The Status of Clustering

Clustering seems to be coming of age as a recognized data-analysis procedure. It is being generally accepted as a means of data compression or summarization, because it automatically places cluster centers in regions of high sample densities so as to represent the full data set by means of a smaller number of prototype, or cluster, points. The success of such a procedure, measured by the error of fit, depends upon the data structure itself. Certain types of data set are therefore more suitable for clustering than others. Because of a range of different types of clustering procedures, and because it is a heuristic rather than an exact method, there are many detailed differences in the clustering methods available.

Related techniques such as clumping, numerical taxonomy, and aggregation are other terms that are used to describe approaches that consider some of the same problems as are faced in clustering. That these techniques are generally coming of age is evidenced by the books of Roach [1], Fisher [2], and Grinker et al. [3]. Many papers on these subjects are reported in the literature, some of which were referenced in a paper by Ball and Friedman [4] on clustering, given at the last annual meeting. A course on clustering is being organized by CEIR, and presented by Ball and Solomon. Of course, the Russians are doing it too (experimenting with clustering), as heard in talks recently given by Professor Aizermann.

There are several batch-processing computer algorithms for clustering that are available on various computers. The ones we know of are our own ISODATA, the Singleton-Kautz program, and the Rubin-Friedman program available through IBM. Thus, by a fairly logical process, if one needs to cluster, a computer must be used, and if some interaction with the program is needed and some graphics output is required for interpretation and control, then some PROMENADE-like system is needed.

c. The General Status of the Current PROMENADE System

We are frequently frustrated by the progress of development of our system, which at times seems to be quite slow. However, in another sense, when we look back we can be partially satisfied with some of the achievements.

Progress is hampered by the following factors:

(1) We seem to be at the forefront of the state of the art in diverse fields. Some of these fields are, clustering and statistical computation, interactive graphics, computer software systems, graphic representations, and the broad field of data analysis.

(2) There are significant basic hardware and software systems problems that must be overcome before we can tackle the problems of our own applications-oriented software.

(3) We require a complex interface and program structure to provide convenience for the user, numerous data-processing methods to match numerous types of data, data-manipulation capabilities, ability to treat input as output, and other features that add complexity to system implementation.

(4) We have had a relatively limited budget for this development.

(5) Upgrading of the computer system from a CDC 3100 to a CDC 3300 has caused problems in the mapping over of system software.

2. General Analysis Procedure

In this section we give an overview of the ways in which the system can be used to analyze data and, more specifically, to perform interactive graphic clustering of multivariate data sets. Later sections of the paper will describe in more detail the features available in each of the subsections.

a. The Users

It is highly desirable, when using the system on new application data, to have at least two persons to conduct the initial explorations. One of these persons should be an expert, or at least knowledgeable, in the use of the system and the features it can provide. The other person should be knowledgeable about the data collected, its manner of collection, and how to relate the results from the system to some useful conclusions in terms of the application problem. This latter function is usually the more difficult to perform. If a third expert can be obtained, we suggest he should have a strong background in applied statistics.

1) Functions of the Expert System User

If only the expert in the use of the system is present, then he is restricted to working with the data in a purposeless manner, without any specific goals in mind. He must then resort to finding features that he considers may be important. For instance, he may note a certain high value of correlation between a particular pair of variables, because this pair might be the only pair that gives a high correlation. He may notice, for instance, that a particular data point appears, both in the plots and the numeric techniques, to be an outlier--i.e., distinctive. He can identify this point and print out its component values. However, his exploration is aimless, and he may be spending time finding out relationships and characteristics that are obvious to anyone who knows about the application. By definition, a useful exploration must discover something that is not anticipated, and usually experts in some application area already know many facts concerning the data relationships.

2) Functions of the Data-Application Expert

If only the expert in the field of the data application is operating the system, he will initially be unfamiliar with the system and will have to learn its use while he is trying to interpret the meaning of the results. Although it is feasible to use the self-teaching features in the system to guide a new user, we do not advise this approach. If it is not possible to have an expert in the use of the system initially to assist the expert in the data's field of application, we strongly advise some preparation on the part of the data expert before he loads the system. This preparation should consist of reading available reports and documentation.

However, we must stress that, from our past experience, much greater progress is made when a knowledgeable system user and an expert in the field of the data can work together at the console at the same time, as illustrated in Figure 1. The system expert should determine how he can use the system to test the hypothesis formulated by the data expert.

b. Typical Analysis Procedure

The operational procedure that will generally be carried out in order to perform interactive clustering will typically involve the following steps. The system must be loaded into the computer memory and suitable disk files must be opened. This operation generally takes a minute or so, and involves loading a disk pack and pressing a series of buttons that causes about 15 cards to be read at the card reader.

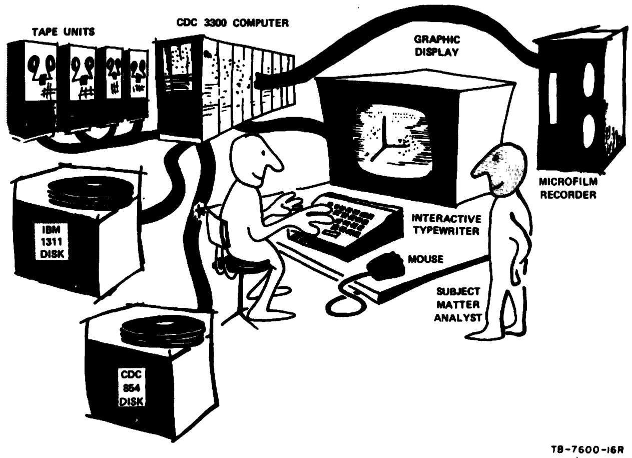


FIGURE 1 TYPICAL MODE OF OPERATION OF THE PROMENADE SYSTEM

The remaining operations can be carried out seated around the graphic console, or going over to the line printer to inspect more detailed output than is available on the display screen.

On the first occasion that a new set of data is read into the system for analysis, a set of data cards must be placed in the card reader and the data on the cards transferred to the disk for more convenient manipulation and semi-permanent storage. Each data file has a name and consists of a two-dimensional data array as shown in Figure 2. The limits on the size of this array

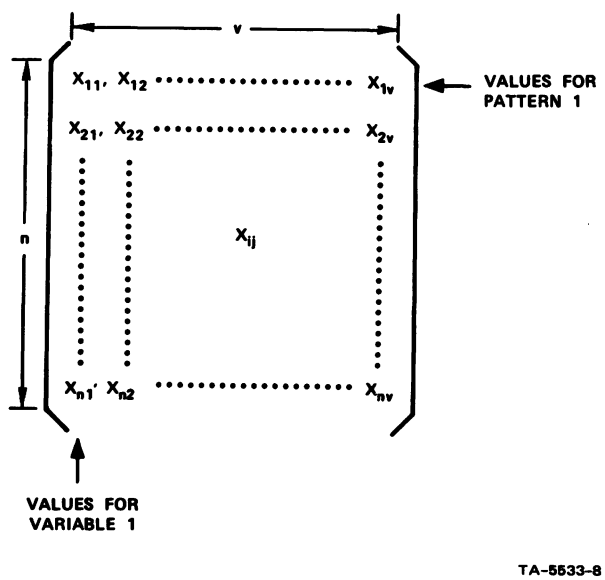


FIGURE 2 DATA ARRAY

are currently fifty variables and one thousand samples. The user of the system chooses which data set he wishes to operate upon by typing in the name of the data set. This data set is then available for manipulation, creation of derived data sets or subsets, or for listing on the line printer together with some overall statistics. In addition, this data set is automatically linked to the graphic and numeric subsystems.

To cluster the data, the user selects the numeric subsystem, and makes a further selection of the ISODATA clustering routine within that subsystem. He may then proceed to cluster the data, using as initial cluster centers either initial cluster centers that he has punched onto data cards, or internally generated initial cluster centers.

After each iteration of the data set, the clustering program returns control to the user, who may guide the clustering process by changing clustering parameters and performing subsequent iterations. The man/machine dialogue occurs through the media of the graphic screen mainly, or else by inspection of printout on the line printer for fine details. The type of information coming out on the screen may be graphical or numeric, while the line printer is used only for numeric and textual information. In graphic form, there are two types of plots associated with the clustering algorithm by automatic means through the file structure. At the end of each iteration, either the link-node plot or the waveform plot representation of the clustering output can be inspected. Furthermore, the clustering can be controlled by specifications on these plots for further clustering iterations.

In this manner, the man/machine dialogue proceeds until satisfactory clustering is obtained or until the user gives up, or gets too tired. Another more practical termination to the clustering session, but one that unfortunately occurs too frequently, is that it becomes somebody else's turn to use the machine. The bad effect of this annoying interruption is minimized, in the system, by storing intermediate results so that continuation can take place from where one left off, without having to start the procedure from the beginning.

The following sections will give more details of the features available in the various subsystems that are used for interactive graphic clustering.

3. Features of the Data-Manipulation Subsystem

We have found that convenient manipulation of experimental data is a valuable feature of an on-line system. For convenience in handling many files of data, each data set is allocated a 20-character short description and an 8-character file name. Each variable may also have an 8-character mnemonic descriptor that is usually printed or displayed on the CRT screen whenever relevant output appears. The variable number (its sequence) is also used on the printout, which makes it not essential to supply mnemonics. Before these labeling features were implemented, the user frequently wasted valuable time identifying his data entities before he could interpret his output.

The patterns, or samples, have similar numerical identifiers, such as sequence and category labels, etc.

There are 10 flags for each pattern as follows:

(1) Sequence number

(2) Category number

(3) ISODATA Cluster number

(4) Rosen-Hall cluster number

(5) ISODATA flag

(6) Rosen-Hall flag (not used yet)

(7) Character for labeling 3D plot

(8)-(10) Not used.

The following options are available to a system user by typing the corresponding keyboard characters. A list of these options is always displayed to the user if he types "O". A fuller discussion of the general command structure that includes global and local options is given in a more detailed report [5]. The page of options is displayed on the CRT to the user when he types O. The page format is 64 columns wide by 32 lines high. The user can also have this page printed out on the line printer, if he wishes, and in this case, column and line numbering, as well as page numbering, is also printed out (See Figure 3). This is the chief mechanism for the system designers to update the options pages. This important function in an evolutionary system is done by punching a deck of cards that can be modified and read in again.

```
0000000001111111111222222222233333333334444444444555555555566666666667777777778
12345678901234567890123456789012345678901234567890123456789012345678901234567890
     DATMAN OPTIONS                                   20 AUG 69    DATMAN   15  1
                                                                   DATMAN   15  2
 C   CREATE A VARIABLE WITH THE ON LINE COMPILER                   DATMAN   15  3
                                                                   DATMAN   15  4
 D   DELETE NAMED FILE FROM THE DISK                               DATMAN   15  5
                                                                   DATMAN   15  6
 E   GO TO EXECUTIVE                                               DATMAN   15  7
                                                                   DATMAN   15  8
 F   SELECT A FILE FROM THE FILE DIRECTORY TO PASS TO ALL THE      DATMAN   15  9
          ROUTINES IN THE PROMENADE SYSTEM                         DATMAN   15 10
                                                                   DATMAN   15 11
 N   READ A NEW DATA SET AND STORE ONTO DISK                       DATMAN   15 12
                                                                   DATMAN   15 13
 P   PRINT THE CURRENT OPENED DATA FILE (PRINTS THE DATA VALUES,   DATMAN   15 14
          MNEMONICS, AND STATISTICS OF A DATA FILE)                DATMAN   15 15
                                                                   DATMAN   15 16
 S   SELECT A SUBSET OF THE CURRENT OPENED FILE AND STORE THE      DATMAN   15 17
          SUBSET IN A NAMED FILE.                                  DATMAN   15 18
                                                                   DATMAN   15 19
 T   FORM A NEW DATA SET BY TRANSFORMING THE DATA ON THE CURRENT   DATMAN   15 20
          OPENED FILE AND STORE IT ON A NAMED FILE.  TWO CHARS     DATMAN   15 21
          REQUIRED.                                                DATMAN   15 22
                                                                   DATMAN   15 23
                                                                   DATMAN   15 24
                                                                   DATMAN   15 25
                                                                   DATMAN   15 26
                                                                   DATMAN   15 27
CURRENT OPENED FILE NAME IS                                        DATMAN   15 28
                                                                   DATMAN   15 29
                                                                   DATMAN   15 30
                                                                   DATMAN   15 31
                                                                   DATMAN   15 32
0000000001111111111222222222233333333334444444444555555555566666666667777777778
12345678901234567890123456789012345678901234567890123456789012345678901234567890
```

TA-5533-9

FIGURE 3 PRINTOUT OF A PAGE OF OPTIONS FOR THE DATA-MANIPULATION SUBSYSTEM

The functions performed by these commands and some additional commands for handling files are next described in more detail. For convenience, the data-manipulation file-handling subsystem is sometimes referred to as "datman."

a. Mode of Operation of Datman Permanent Files

There are 25 permanent files available to the PROMENADE System, with five being occupied by certain data sets used for verifying and exercising the system. These files may not be deleted or overwritten. The remaining 20 are available to System users.

Only one data file will be open and available to all the routines in the System at one time. The file, which is opened by automatic default when first starting the system, is the KENDALL file.

The following commands change the opened file:

F Select a file to use in the System. The file directory is displayed and the operator is required to type in a file name. This name is compared with the names in the file directory. If it matches one of them, the file it matches is opened and the other one is closed. If no match is found, the old open file remains open.

S Select a Subset. The file directory is displayed and the operator is required to type in a new file name. This new file name must be different from the current opened file. If it matches the name of a different file, that file will be overwritten. The operator may then proceed to select a subset of his data set by selecting a subset of the variables. Note that the Command Delete (CD) key on the keyboard deletes the last variable. After selecting variables, the operator selects the indices of a DO loop to select a subset of the patterns.

N Read a new set of data from cards. A deck of data is read from the card reader and stored on the disk. The first card in the data deck contains the file name--i.e., data set name. This card is placed in front of the run-time format card. The file name for the data is compared with the file names in the file directory. If it matches a name in the directory, that file is overwritten with the data. If no match is found, the data set is put on a blank file. If there are no blank files, the operator must delete a file before reading the data.

T Transform the data. This works much the same as the select-a-subset option. It displays the file directory and allows the operator to type in a file name under which to put his transformed data. This name must be different from the name of the currently opened file. The name is compared to the names in the file directory. If it matches the name of a file, that file will be overwritten. The operator then must select the variables to be transformed. This is done by pointing at the variable mnemonics on the screen with the mouse and pressing the button on the mouse to select the variable. Typing Command Delete (CD) deletes the last variable that was selected. Type "space" as soon as the list of variables is complete. The operator then must select a transformation by typing one of the following characters:

A Reciprocal of square root $X_{new} = 1/\sqrt{X_{old}}$

E Exponential $X_{new} = e^{Xold}$

L Logarithmic $X_{new} = Log_e X_{old}$

M Fill in missing data--type in a value. All data found equal to it will be replaced with the average for that variable (the average is calculated with the value and all like it deleted).

N Normalize the data--mean to 0.0 and standard deviation to 1.0

Q Square $X_{new} = (X_{old})^2$

R Reciprocal $X_{new} = 1/Xold$

S Square root $X_{new} = \sqrt{X_{old}}$.

At the end of the transformation the new file containing the transformed data is left open and the file containing the untransformed data is closed.

C Create a new variable. This allows the user to derive new variables from any combination of existing ones, using arithmetic and functional forms as well. After typing C, the user must enter the name of a new variable. This name must be in the format of a standard FORTRAN identifier. Then must follow a relational expression using at least one already defined variable, and any other already defined variables that the user may wish. For example:

```
NEWVAR3 = PETLEN * 2.15
+ (SEPLEN ** 2)/9.866
```

would be a valid expression. This extra variable would be automatically added as an extra column to the data file, after the necessary calculations had been carried out for each row of the data set. The new expression is also printed out on the line printer as a permanent record.

The following command may change the opened file:

D Delete a file. This command allows the operator to delete a file from the disk. The file directory is displayed. The operator must type in the name of the file he wants to delete. The name is compared with the file names in the file directory. If it matches the name of a file in the directory, the file is deleted from the disk. If it matches the name of the currently opened file, the file is deleted and the KENDALL file is left open. If no match is found, a message is printed and nothing happens in the program.

The following commands cause no change in the opened file:

P Print the data. This command prints the file header, data statistics, and data values of the currently opened data file.

E Exit. This command returns control to the Executive Subsystem of the PROMENADE System.

These data-manipulation features mainly provide convenience to the user. Similar functions can be performed off-line, or by re-punching cards, etc., but only at the cost of a much greater delay. In fact, the difference in delay is so many orders of magnitude that the ability to perform these manipulations on-line and then immediately use the results is a qualitative difference rather than a quantitative one in effect. If the manipulations were not so convenient to do, they probably would not get done by the other, slower means.

4. Features of the Numerical Clustering Algorithms

The clustering algorithms presently integrated into the system are the ISODATA clustering algorithm and the Rosen-Hall pattern-recognition algorithm. This latter performs a clustering, subject to the constraint that all patterns composing a cluster must have the same category label. The ISODATA algorithm, due to recent improvements to the system, can be controlled from three different points as shown in Figure 4.

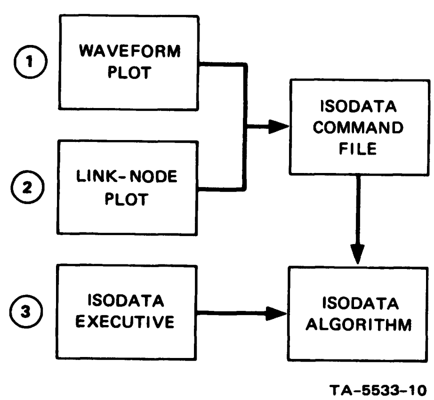


FIGURE 4 THREE CONTROL MODES FOR ISODATA

The controls from the two plots are treated in more detail in the following section. Briefly, they allow the user to watch the progress and control the clustering while viewing the plots, rather than having to go backward and forward from the ISODATA executive to the plots--a process that takes about 10 seconds and is annoying to perform between each iteration.

The ISODATA executive allows for overall control of the algorithm, for parameter changes, printout controls, etc., as fully documented in other reports [5]. The clustering performs grouping of data points that are close together in multivariate space. This data-compression or fitting process is useful when trying to summarize data. Although there are many types of summary statistics and graphs that are produced as output from the clustering, perhaps the most global information is obtained from the clustering characteristic curve, as shown in Figure 5.

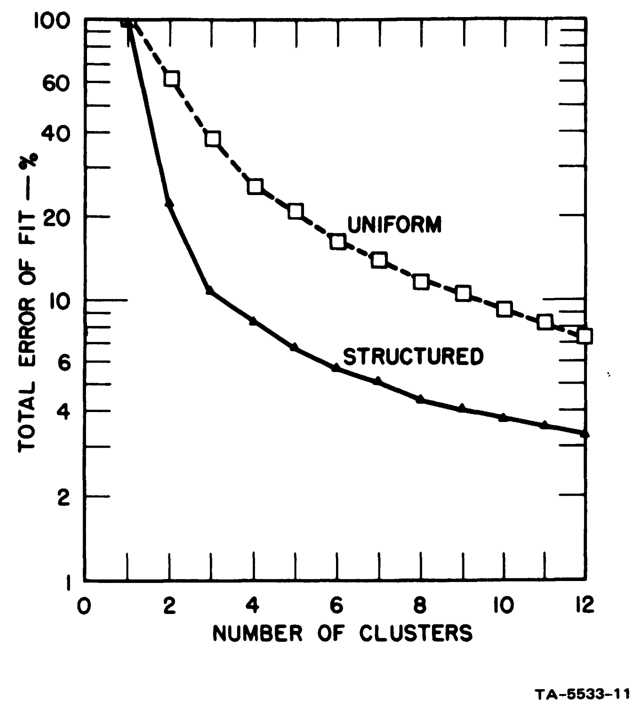
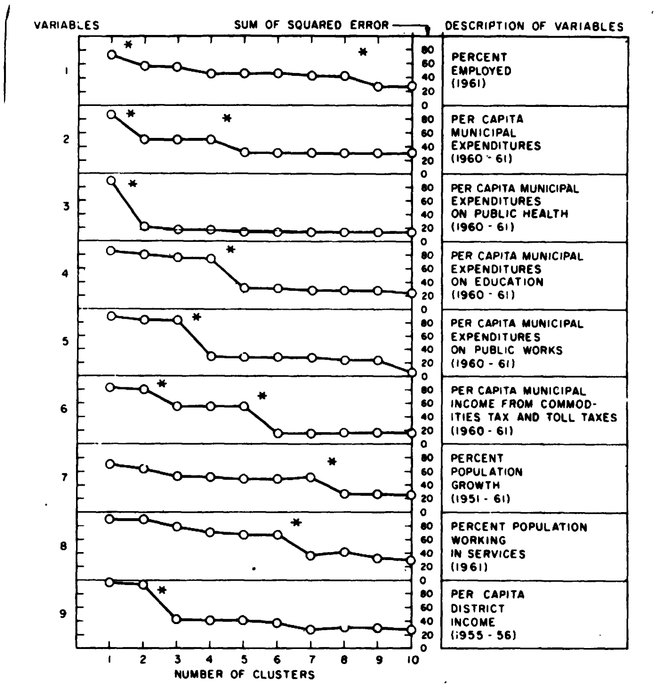


FIGURE 5 CLUSTERING CHARACTERISTICS FOR UNIFORM AND TYPICAL STRUCTURAL DATA

The way in which each individual variable contributes to the total error of fit can also provide useful information, as shown in Figure 6. This is a convenient way of showing which variables take part in the partitioning of clusters, and at what stage of the process. Other information that is useful appears on the display screen in textual and numeric form. For example, the iteration number, number of clusters, parameters and printout values, etc. are displayed. More detailed results appear on the line printer, such as the positions of all the cluster centers, and which patterns they contain, how many are in each cluster, and what the average distance (dispersion) is in each cluster. The user can choose plotting options from the ISODATA executive, either the link-node or waveform plots, to view the results. Linkage of the clustering output data as input to the plots is done automatically by the system. To "go" to one of the plots, the user must type "E", to return to the executive of the system, and then type "L" for link-node (see Figure 7), or "W" for waveform (see Figure 8).



TA-5533-12

FIGURE 6 CLUSTERING CHARACTERISTICS FOR INDIVIDUAL VARIABLES

The ISODATA algorithm has been steadily developed over a period of about five years, from a batch program to an on-line interactive version with associated graphics. Both ALGOL and FORTRAN versions for many different machines have been developed. As clustering continues to be useful in many applications we find that we can provide new features to give greater insight and control, and to carry out the clustering more rapidly, with greater conveneience.

The user can review the options available to him by typing 0, and in this case, the options as given in Figure 9 will be displayed. The IPRINT parameter is displayed on the parameter control page. If the value of IPRINT is changed from its initial value of zero, fairly extensive printout will be obtained on the printer. If the value of IPRINT is set to 11, then, in addition to the normal printout, the distance of each pattern from its cluster center is also printed. In order to accommodate a wide range of magnitudes of numbers, because this is designed to be a general-purpose system, an internal calculation is done on numbers before they are printed. Depending on the results of this calculation, an appropriate run-time output format is chosen. This avoids the cumbersome form of output sometimes seen used in other systems where 8995.2600000E-3 is used instead of 8.99526.

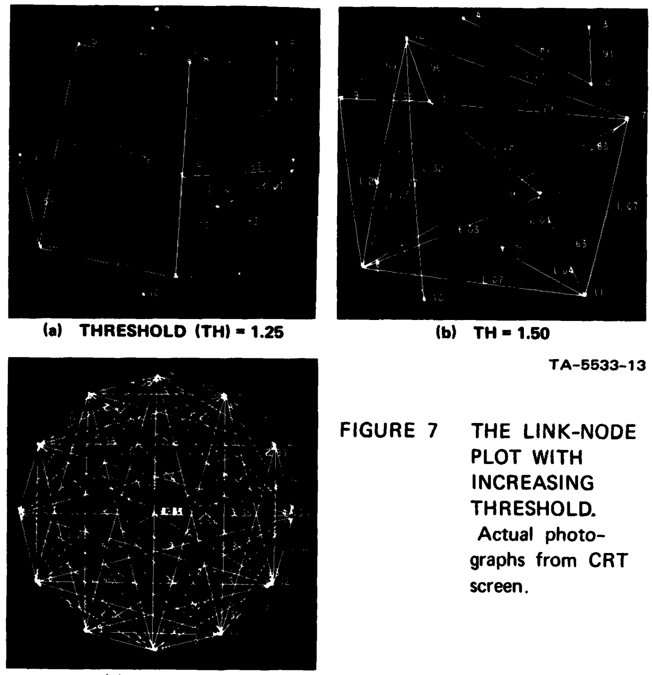
5. Graphic Control of Clustering

a. Graphic Features in PROMENADE

Graphical representation of multivariate data is achieved in the system through several types of plots. These plots usually provide automatic scaling and labeling features, automatic linking to data files, and can generally plot data faster than a human can absorb the meaning of the plots. This level of automation is a significant advantage over other techniques. The interactive nature of the plots, and the ability to hunt for a good view, allows for selection and capture, in hard-copy form, of only significant information. In the batch approach, a much larger volume of hard copy information usually has to be captured in order to find the significant information. Providing the user with effective graphic insight to multivariate data is quite a challenge to the display programmer. It is also evident that a single type of graphical transformation will not be adequate for general data-analysis purposes. Hence, in PROMENADE, numerous plots are available. Currently, the following plots are provided:

(1) Scatterplot, or two-dimensional plot
(2) Multivariate histogram plot
(3) Distance-along-versus-away-from a line plot
(4) Psuedo three-dimensional plot
(5) Pseudo four-dimensional plot
(6) Waveform plot, or profile plot
(7) Link-node plot

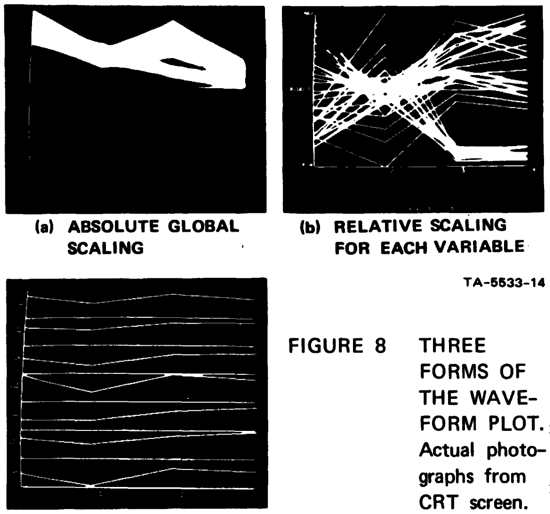
These plots are mostly of a dynamic nature, and each can be controlled by numerous control parameters. For example, the pseudo four-dimensional plot allows the user to get a true perspective view of his data as though he was flying about in it. Although all these plots can aid exploration of a data set that is to be clustered, and thus indirectly aid the purposes of clustering, direct control of the ISODATA clustering has only been implemented via the latter two plots.



(a) THRESHOLD (TH) = 1.25 (b) TH = 1.50 (c) TH = 9.00

TA-5533-13

FIGURE 7 THE LINK-NODE PLOT WITH INCREASING THRESHOLD. Actual photographs from CRT screen.



(a) ABSOLUTE GLOBAL SCALING

(b) RELATIVE SCALING FOR EACH VARIABLE

(c) INDIVIDUAL DATA POINTS

TA-5533-14

FIGURE 8 THREE FORMS OF THE WAVEFORM PLOT. Actual photographs from CRT screen.

b. Graphic Control Through Both Plots

Examples of both plots are shown in Figures 7 and 8. Each plot has a set of options or parameters that can be changed to change the plots themselves. In addition they have commands for controlling the ISODATA clustering. These latter commands are identical, but each plot has its own unique plot parameters for controlling the data representation. For example, Figure 10 shows a printout of the page of options for the waveform plot. This page appears on the CRT screen when the user types "O".

```
00000000011111111112222222222333333333344444444445555555555666666666677777777778
12345678901234567890123456789012345678901234567890123456789012345678901234567890
AT THE BEGINNING OF EACH ITERATION, YOU MAY CHANGE THE VALUES OF ISODATA  23  1
THE LAST 7 PARAMETERS SHOWN ON THE PARAMETER DISPLAY.  CHOOSE A  ISODATA  23  2
NUMERIC PARAMETER WITH THE MOUSE AND TYPE IN THE NEW VALUE.      ISODATA  23  3
THE LAST TWO DIGITS TYPED FOR THETAC WILL BE TAKEN AS THE        ISODATA  23  4
FRACTIONAL PART.  OTHER OPTIONS ARE...                           ISODATA  23  5
                                                                 ISODATA  23  6
I        REINTRODUCE INDIVIDUAL PATTERNS INTO THE CLUSTERING SET ISODATA  23  7
D        DELETE INDIVIDUAL PATTERNS FROM THE CLUSTERING SET.     ISODATA  23  8
X        PUNCH OUT THE CLUSTER CENTERS AND THE PCM TABLE.        ISODATA  23  9
E        RETURN TO THE EXECUTIVE.                                ISODATA  23 10
O        DISPLAY THE LIST OF OPTIONS.                            ISODATA  23 11
                                                                 ISODATA  23 12
COMPUTATIONS ARE BEGUN BY TYPING ONE OF THE FOLLOWING VALUES FOR ISODATA  23 13
THE SPLIT/LUMP PARAMETER...                                      ISODATA  23 14
                                                                 ISODATA  23 15
S        SPLIT...A TRIAL SPLITTING OF CLUSTERS IN TWO IS         ISODATA  23 16
              PERFORMED AND THE MEANS OF THE SUB-CLUSTERS        ISODATA  23 17
              COMPUTED.  IF THE MEANS ARE MORE THAN 1.1THETAC    ISODATA  23 18
              UNITS APART, THE NEW CLUSTERS ARE KEPT AND THE OLD ISODATA  23 19
              ONE DELETED.                                       ISODATA  23 20
L        LUMP...TWO CLUSTERS ARE COMBINED IF THE MEANS ARE LESS  ISODATA  23 21
              THAN OR EQUAL TO THETAC UNITS APART.  AT MOST      ISODATA  23 22
              NCLST PAIRS OF CLUSTERS WILL BE LUMPED IN AN       ISODATA  23 23
              ITERATION.                                         ISODATA  23 24
N        NEITHER SPLIT NOR LUMP.                                 ISODATA  23 25
A        THE PROGRAM WILL DECIDE WHETHER TO SPLIT OR LUMP.       ISODATA  23 26
SPACE    USE THE PREVIOUS VALUE OF THE SPLIT/LUMP PARAMETER.     ISODATA  23 27
                                                                 ISODATA  23 28
CLUSTERS WITH FEWER THAN THETAN PATTERNS ARE DELETED.            ISODATA  23 29
FOR NO LINE PRINTER OUTPUT, SET #IPRINT# TO ZERO.                ISODATA  23 30
                                                                 ISODATA  23 31
PRESS THE SPACE BAR TO REDISPLAY THE PARAMETERS.                 ISODATA  23 32
00000000011111111112222222222333333333344444444445555555555666666666677777777778
12345678901234567890123456789012345678901234567890123456789012345678901234567890
```

TA-5533-15

FIGURE 9 COMMAND OPTIONS FOR ISODATA

A typical command structure for controlling ISODATA from this plot can be written in general form as:

XY (parameter string) (command accept)

where

XY is one of the listed two-letter commands

(parameter string) is a list of cluster or pattern numbers separated by commas

(command accept) is a special keyboard character.

For example, the command:

KS 5,7,14 (command accept)

would cause splitting of the clusters 5, 7, and 14.

c. Interpretation Using the Waveform Plot

The waveform plot is more familiar to psychologists as the profile plot. In Figure 7 examples of this plot in its various forms is given. The figures show a four-dimensional data set. In Figure 8(a) the scaling is global--i.e., the same scale is used for all dimensions and the data is placed symmetrically about a zero-valued center line. The maximum data value is displayed at the maximum height of the vertical scale. The number of samples on the screen/total number of samples in the data set, is also given. Provisions for selecting subsets of data are made in this subsystem. Figure 8(b) is similar except for the scaling, which is local to each dimension. This allows a more detailed examination of the range of data values in each dimension but does not allow magnitude comparisons from one dimension to the other. The scaling is arranged to place the maximum and minimum values at the extremities of the vertical axis. The ability to visually observe clustering of data, and to see which dimensions are contributing to clustering, is the chief advantage of these superimposed waveform plots.

In Figure 8(c) the data is displayed by individual data point, and numbered sequentially on the left to allow identification. Scaling can be either global or local in this plot also. When the user depresses the space bar, the next group of seven data points is shown. Cluster centers can also be displayed in these same formats since they are prototype data points, but they are displayed using dotted lines to distinquish them.

d. Interpretation Using the Link-Node Plot

In the current version of the link-node plot only the cluster centers are displayed. These are initially placed on the circumference of a circle, as shown in Figure 7(c). Initially no links are displayed--only nodes and their corresponding numbers. This is a non-geometric plot in that multivariate data values are arbitrarily placed initially, and in that many dimensions are "squashed" down to two. A threshold value can be introduced to create links. Links are drawn between any two nodes if their distance apart in the original multivariate space is less than the value of threshold used. This plot can also be described as a way to graphically represent a two-way table of distances. Since the CRT screen is a volatile dynamic medium, it is easier to perform these operations in this way rather than by pencil and paper.

In Figure 7(a), the initial positions of the nodes have been moved, and the threshold value (TH) is set to 1.25 by the operator. Note that there are two connected networks of points, and one isolated cluster center, number 10, at this threshold value. Raising the threshold value to 1.50, as shown in Figure 7(b), causes more links to be drawn. This figure also illustrates some rearrangement of node positions by the operator. Now cluster center 10 is linked, but the two networks are still separated. In Figure 7(c), all possible pairs of links are drawn with a threshold value of 9.00.

6. Applications

Applications can be divided into two groups:

(1) A small application, involving typically a few days of discussion and interpretation, and a few hours of computer time.

(2) Significant applications, involving long-term contractual agreements and substantial amounts of effort and money.

In Category 1, we have processed the following data:

- Agricultural data for pineapple growth
- Economic factors for Indian cities
- Nutritional data for Hawaiian school children
- Socio-economic factors for gasoline station placement, etc.

In Category 2, four applications are in progress:

- Chemical, for classification and clustering of spectrometer data
- Marketing, to classify companies from their public financial data
- Census, comparison of regression and clustering methods for housing-construction price index
- Cloud clustering from satellite photographs.

It is not the purpose of this paper to discuss applications of interactive clustering in detail. A paper will be presented discussing the four applications in Cateogry 2 in detail [6].

7. Conclusions

We have a significant facility for performing interactive graphic clustering. The last four months have been dedicated to changing over to the new computer. We hope the next phase will be one of active use of the system on significant applications. Meeting the needs of the application problems is the ultimate function of the system, and many ideas for system improvement are generally gained in trying to better meet the application needs.

```
00000000011111111112222222222333333333344444444445555555555666666666677777777778
12345678901234567890123456789012345678901234567890123456789012345678901234567890
    OPTIONS FOR WAVEFORM PLOT                          20 AUG 69  WAVEFORM 33  1
                                                                  WAVEFORM 33  2
                                                                  WAVEFORM 33  3
    G    PERMUTE VARIABLES                                        WAVEFORM 33  4
    P    SELECT POINTS FOR THE PLOT                               WAVEFORM 33  5
    O    DISPLAY THIS PAGE OF OPTIONS                             WAVEFORM 33  6
    M    CHANGE MODE OF PLOT                                      WAVEFORM 33  7
    E    EXIT                                                     WAVEFORM 33  8
    S    CHANGE SCALE FOR PLOTS                                   WAVEFORM 33  9
                                                                  WAVEFORM 33 10
                                                                  WAVEFORM 33 11
                                                                  WAVEFORM 33 12
                                                                  WAVEFORM 33 13
                                                                  WAVEFORM 33 14
         THE FOLLOWING COMMANDS CONTROL THE ISODATA CLUSTER       WAVEFORM 33 15
                   ANALYSIS ALGORITHM.                            WAVEFORM 33 16
                                                                  WAVEFORM 33 17
    KB   BEGIN ISODATA CLUSTERING -- TYPE IN SPHERE FACTOR        WAVEFORM 33 18
    KC   CREATE A CLUSTER CENTER                               NI WAVEFORM 33 19
    KD   DELETE THE SPECIFIED CLUSTER(S)                          WAVEFORM 33 20
    KF   FREEZE THE SPECIFIED CLUSTER(S)                          WAVEFORM 33 21
    KG   GO TO ISODATA AND DO SOME SETTLING ITERATIONS            WAVEFORM 33 22
    KL   LUMP THE SPECIFIED CLUSTERS                              WAVEFORM 33 23
    KS   SPLIT THE SPECIFIED CLUSTER(S)                           WAVEFORM 33 24
    KU   USE THE SPECIFIED CLUSTER(S)                             WAVEFORM 33 25
    PD   DELETE THE SPECIFIED PATTERN(S)                       NI WAVEFORM 33 26
    PF   FREEZE THE SPECIFIED PATTERN(S)                       NI WAVEFORM 33 27
    PU   USE THE SPECIFIED PATTERN(S)                          NI WAVEFORM 33 28
    VI   IGNORE THE SPECIFIED VARIABLE(S)                      NI WAVEFORM 33 29
    VU   USE THE SPECIFIED VARIABLE(S)                         NI WAVEFORM 33 30
                                                                  WAVEFORM 33 31
                                        NI = NOT IMPLEMENTED      WAVEFORM 33 32
00000000011111111112222222222333333333344444444445555555555666666666677777777778
12345678901234567890123456789012345678901234567890123456789012345678901234567890
```

FIGURE 10 OPTIONS FOR WAVEFORM PLOT

References

1. Roach, S. A., "The Theory of Random Clumping," Methuen and Co., London (1968), pp. 94.

2. Fisher, W. D., Clustering and Aggregation in Economics, Johns Hopkins Press, Baltimore, Maryland (1969).

3. Grinker et al., The Borderline Syndrome, Basic books (1968). See especially Chapter 5.

4. Ball, G. H. and Friedman, H. P., "On the Status of Applications of Clustering Techniques to Behavioral Sciences Data" Proceedings of the Social Stats. Sec. (1968), pp. 34-39.

5. Hall, D. J. et al., "Promenade--An Improved Interactive-Graphics Man/Machine System for Pattern Recognition," Final Technical Report, RADC-TR-68-572, Rome Air Development Center (June 1969).

6. Hall, D. J. et al., "Applications of the PROMENADE Data-Analysis System," submitted to Computers and Communications Conference, Mohawk Valley Section IEEE, Rome N. Y., 1 October 1969.

RECENT DEVELOPMENTS IN SEQUENTIAL DATA ANALYSIS STRATEGY

John A. Sonquist, University of Michigan

Abstract

Within the past few years computerized sequential algorithms have been developed and used to search out and display non-additive relationships implicit in survey data. The objective is the display of information relevant to the problem of how to specify a multiple classification analysis model purporting to explain the phenomenon in question. Experimentation with one such algorithm (AID) has led to a classification of simple additive and interactive multivariate models related to elementary Boolean operators. These models are described and illustrated; extensions to the AID algorithm to facilitate specification of complex models, for dealing with covariate models arising from "crucial" variables or over-time survey data and for imposing symmetry restrictions are discussed. Preliminary experimentation with the improved algorithm is reported.

Keywords: Survey, Multivariate analysis, model specification, statistical interaction, computer program, prediction, analysis of variance, multiple regression, multiple classification analysis, data analysis strategy, simulation.

The task of interpreting a mass of non-experimental data, such as that generated by modern cross-sectional and longitudinal survey methods has remained a difficult one. Two reasons for this appear to stand out as particularly worthy of examination. One is that the statistical methods used by most data analysts are oriented toward deductive testing of isolated hypotheses rather than toward the more frequent task of extraction of information from data and the development of models on an inductive basis. In addition, the full power of "second generation" computing equipment has not yet been utilized (let alone "third generation"); it has the ability to extend the scope of the analyst's logic as well as perform computations rapidly for him.

The present paper is a modest attempt to surmount these obstacles by extending the idea of coupling the calculating power of modern computing equipment to a formal sequential algorithm, the objective of which is the generation of information about the data. In particular, our concern is with obtaining summary information organized for use by the analyst in the inductive stages of his research; that is, in the stages of model formulation and development. The objective of the procedure is the display of information relevant to the problem of how to formulate or specify a model purporting to explain the phenomenon in question.

One frequently appearing and concrete form taken by this general specification problem is the question of choosing and of specifying the form of the terms to be included in a least squares multiple classification equation (MCA). Such an equation is commonly used as a statistical model representing the simultaneous and direct explanatory effects of a correlated set of presumed causal factors on a single dependent variable. It is of the form

$$Y_{ijk...} = \bar{Y} + a_i + b_j + \dots e_{ijk...} \quad (1)$$

where

- Y_{ijk} = the score (on the dependent variable) of individual k, who falls in category i of predictor A and into category j of predictor B, etc.;
- $\bar{Y}$ = the grand mean of the dependent variable;
- a_i = the "effect" of membership in the i-th category of predictor A expressed as a deviation from the grand mean and adjusted for the intercorrelations of predictor A with the other terms in the equation;
- b_j = the corresponding "effect" associated with predictor B;
- $e_{ijk...}$ = the error term for the k-th individual.

Any given term in the equation, say b_j, may represent the effect of a two-way or higher-order interaction term between explanatory factors not otherwise represented in the equation.[1]

[1] For a more detailed exposition of this basic technique, see Andrews, Morgan and Sonquist (1967). A simple n-way analysis of variance model is not used because the number of observations in the cells of an n-way cross-classification of the predictors are not equal or proportional. This condition ordinarily occurs in survey or other non-experimental data in which there are non-zero correlations between explanatory characteristics.

It is this problem of choosing terms to be included in the equation to which we address ourselves. We shall review some of the previous work in this area, outline several improvements in an algorithm proposed several years ago, and present an illustration of the improved technique.

Previous Work in This Area

In a previous paper before this group, two of the present authors reported preliminary results from a new method of data analysis. We had undertaken to design, program, test, and document a large scale computer program. The algorithm incorporated into the program was aimed at extracting information bearing on the need to introduce interaction terms into a multivariate analysis in which a set of correlated predictor variables were to be related simultaneously to a criterion. Starting from a consideration of some of the problems inherent in applying multivariate statistical techniques (particularly multiple regression) to cross-sectional survey data, we had concluded that the use of Multiple Classification Analysis (MCA) permitted adequate handling of most data characterized by intercorrelated predictors, nominal scales and non-linearities, but could not deal with interaction effects. The computerized procedure that was developed at that time attacked the problem of locating interaction terms by asking a different kind of statistical question of the data than is implied by the immediate choice of MCA.[1]

The technique (termed Automatic Interaction Detection [AID]) employs a sequential decision procedure to divide the sample into a mutually exclusive set of sub-groups through a series of successive binary partitions, each formed by combining observations which are alike on certain characteristics. At each stage every observation is a member of exactly one such sub-group. These groups are formed so that at each stage in the process their means account for more of the total sum of squares (i.e., reduce the error sum of squares more) than the means of any other pair of such sub-groups obtainable by the algorithm. Thus, at each stage in the branching process, the set of groups developed at that point represents, according to the criteria of the model, the best currently available scheme for predicting the dependent variable from the information available. The branching process terminates when all existing sub-groups are either so homogeneous that the desired accuracy has been achieved, or no variable can be found which will enable an improvement in prediction sufficient to warrant its use.[2]

An examination of the statistics computed during the tree-like partitioning process provides evidence in support of, or against, the introduction of additivity assumptions, as well as providing some indication of what kind of interaction term should be generated, if required.

Experimentation with the method on data which were constructed to violate additivity assumptions revealed that asymmetric tree structures were associated with the existence of predictors which interacted.[3] On the other hand, additivity was associated with symmetric structures.

Interactive data were also found to be associated with the presence of differential changes in explanatory power displayed by corresponding predictors in different branches of the tree. Additivity was associated with similar changes in explanatory power.

Inspection of the profiles of mean values defining the effects of a given predictor in various branches of the tree also provided evidence of the appropriateness of making additivity assumptions. Similarity of effect profiles of a predictor in various parts of the tree structure implies additivity; but interactive data produce incongruent profiles.

AID was first proposed as a substitute for MCA. However, experience with the method in actual use and knowledge gained through experimentation with known data structures led to the development and publication of a strategy for using the AID and MCA techniques

---

1 Preliminary results are reported in Morgan and Sonquist (1963b).

2 Details of the process, flow charts, the computer program, sample input and output, formulas, nine illustrative analyses and recommendations for interpreting the output of the program are reported in Sonquist and Morgan (1964).

3 This experimentation is reported in Sonquist (1969a).

jointly, to supplement each other.[1] In addition it appeared that substantial gaines in analytic power might be made from implementing a somewhat more sophisticated sequential algorithm.[2]

Since its original implementation in a large scale program (Sonquist and Morgan, 1964) the algorithm has been translated by others to run on a number of other computers (Campbell, 1965; Aptakin, 1965; Land, 1965; Kay, 1966; Marks, 1966). Versions of the AID algorithm were also adapted to European equipment (Biervert, 1966, Arpi, 1967). The methods have been used by economists (e.g., Snowbarger, 1967; Gensemer, Lean and Neenan, 1965), by sociologists (Ross and Bang, 1965), by political scientists, (e.g., Sarlvik, 1968), by marketing researchers (e.g., Arpi, 1967), by engineers (e.g., Carlson, 1967), and by psychologists, (Caplan et al, 1966).

Revisions in the Algorithm

Experience in using the original algorithm over a period of five years, discussions with others who have used it, and experimentation with contrived data have all led to a rather large collection of proposed improvements in the techniques. Most of these are minor from an analytic standpoint, such as obtaining output in improved form.[3] They are obtained simply as a byproduct of the reprogramming necessary to incorporate the major changes proposed below.

The major revisions include extending the original algorithm to deal with covariance as well as a profile of means, providing for the imposition of an optional premium for symmetric partitioning, and extending the sophistication of the search algorithm to include a "look-ahead" to several successive partitions. In the latter case, the mechanism would not proceed blindly, always trying to maximize the explained variation in the current partition, but, like a chess player, would explore the possibility of sacrificing present payoff in favor of even greater gains from subsequent "moves." It would also attempt to provide the simplicity and parsimony of a symmetric model where the data appeared to warrant its use. Each of the revisions is discussed below in some detail.

Covariance Search Routine

There are many situations in economic, sociological and psychological research where a multivariate analysis is required, but where there exists one dominant explanatory or control variable. The explanatory factor in question may be of particular importance to some theoretical edifice, it may be subject to control more easily than other factors, or it may simply have been measured much more readily and reliably than the others. Where the data come from an experiment and not a survey, the obvious procedure is covariance analysis.

However, with non-orthogonal survey data, one may want to search out subgroups in which there are different relationships between the dependent variable and this dominant explanatory variable. For instance, in much analysis of cross-section survey data, the economist is often interested in the effect of personal or family income on some behavioral variable, and on whether that "income effect" (as represented by the slope of the regression of the behavioral variable on income) varies with other circumstances. The answer to this question will help to determine whether it is necessary to disaggregate the data in models used for forecasting, and the optimal way to do it.

Sociologists and psychologists often face similar problems in which the purpose of the investigation requires isolating the effect of a particular variable under a wide variety of combinations of circumstances. For instance, intelligence, alienation and authoritarianism have each been the subject of repeated investigations in which the object has been to relate that particular factor to specific consequences in such a way as to specify the form of the relationship under various conditions and for particular types of people.

Another illustration is in the analysis of changes taking place over time.

---

1 The strategy developed was suggested in Andrews, Morgan and Sonquist (1967), and then elaborated further by Sonquist (1969a, 1969b).

2 See Sonquist (1967).

3 Some of the programming revisions might include the use of Fortran IV for compatibility with various manufacturers' equipment; a trichotomous partition option; improved input flexibility with regard to both data and control language; additional output including predictor summary tables; and improved variable transformation capabilities.

The initial value of a phenomenon under study clearly affects its value measured at a subsequent time. This is why the residuals from the regression of its current (t_2) value on its initial (t_1) value are often used as a measure of change, instead of the raw increments.[1] However, this "initial value" effect might not be the same for all sub-groups in the population. If, then, a single equation were to be fitted, a downward bias would be exerted on the correlations between change and those factors thought to be responsible for it. Thus, when residualizing a variable for study, a search should be made to determine if this initial value effect is homogeneous throughout the population.

For all these reasons, a variant of the earlier sequential data analysis algorithm has been developed in which the criterion for sequential subdivision of a sample is changed from the sum of squares explained by two sub-group means (instead of one pooled mean) to the sum of squares explained by two simple sub-group regressions (instead of one simple regression of the pooled data). The search algorithm otherwise has basically the same framework: each group potentially to be divided is examined using all feasible partitions based on each explanatory characteristic.[2] The difference is that the quantity maximized when a group is divided into sub-groups is the variation explained by the regression of the dependent variable on the covariate within each sub-group. If the two best fitting regression lines differ as to intercept or slope, then the unexplained variation would be reduced, and the split with the largest difference between the two regressions would reduce it the most.

Given this conceptualization of the problem, a partition may be chosen so as to maximize any one of three quantities. One may evaluate reductions in unexplained variation due to differences between means, differences in regression

[1] For a thorough discussion of this problem see Lord (1967).

[2] If the order of the (k) classes of the explanatory characteristic is maintained, there are only (k-1) ways of forming two groups on the basis of that predictor. However, if one re-orders them on the basis of the means of the dependent variable there are many more possible ways of forming the two sub-groups.

lines, or both taken together. The original AID algorithm sought to maximize the sum of squares explained by means of the sub-groups resulting from the partition of the "parent" group, i.e., it maximized the expression

$$N_1\bar{Y}_1^2 + N_2\bar{Y}_2^2 - N\bar{Y}^2 \qquad (2)$$

The rationale for this is easily seen in Table 1.

It can be seen from Table 2 that a second quantity could also be maximized.[1] This is the expression

$$(N_1-1)r_1^2 s_{y1}^2 + (N_2-1)r_2^2 s_{y2}^2 - (N-1)r^2 s_y^2 \qquad (3)$$

This is the sum of squares of the two regression estimates around the two group means resultinf from the partition.[2] It reduces to:

$$\sum_{i=1}^{2}\left[\frac{[\Sigma(Y-\bar{Y}_i)\ (X-\bar{X}_i)]^2}{\Sigma(X-\bar{X}_i)^2}\right] - \frac{[\Sigma(Y-\bar{Y})(X-\bar{X})]^2}{\Sigma(X-\bar{X})^2} \qquad (4)$$

A third quantity which may be maximized is obtained from the sum of these two and would represent the effect of both the variable used in the partition and the X covariate. The variation explained by the sub-group means and that explained by the regression are both maximized. This expression is

$$N_1\bar{Y}_1^2 + (N_1-1)r_1^2 s_{y1}^2 + N_2\bar{Y}_2^2 + (N_2-1)r_2^2 s_{y2}^2 - N\bar{Y}^2 - (N-1)r^2 s_y^2 \qquad (5)$$

The existing algorithm has also been modified to permit an examination of the effects of one crucial categorical predictor in various parts of the sample without permitting it to be used in the partitioning process. Permitting the analyst to retain the ability to conceptualize the effects of the crucial variable in terms of slopes and intercepts as well as the ability to use profiles of sub-group means gives a desirable simplicity.

[1] Computational formulas are, of course, somewhat different and are not given here. In maximizing any of these expressions the last term is constant over all possible partitions and can be ignored. See Walker and Law (1953) pp 210-216.

[2] See Walker and Law (1953), pp 242-244.

TABLE 1.

Analysis of Variance for Differences in Means

Source of Variation	d. f.	Sum of Squares	Mean Square
Observations around grand mean	N-1	$\sum_{j=1}^{k}\sum_{i=1}^{N_j}(Y_{ij}-\bar{Y})^2 = SST$	s_y^2
Between group means	k-1	$\sum_{j=1}^{k}N_j(\bar{Y}_j-\bar{Y})^2 = SSB$	MSB
Within Groups	N-k	$\sum_{j=1}^{k}\sum_{i=1}^{N_j}(Y_{ij}-\bar{Y}_j)^2 = SSW$	MSW

The covariate problem is illustrated below. Owning a home (as opposed to renting, etc.) is related not only to income, but to age, family size, and place of residence (urbanization). More important, a very large fraction of older people outside the large urban areas own a home regardless of their income; i.e., income differences have no effect at all on home ownership in this group. However, among young families with children, small increases in income appear to lead to substantial in creases in the probability that the family will soon own its home. Furthermore, among young people with no children at all, home ownership is rare at any income level. Clearly, economic projections of home ownership need to be based not just on aggregate statistics of income increases, but on who received them.

The effective use of survey data to shed light on such problems requires concentrating attention on the differential relation of income to economic activity as well as studying its effect in the aggregate.

These details are further illustrated by an examination of Figure 1. In the total sample the regression of Y on X_1 is

$$Y = a_1 + b_1X_1 + u_1 \qquad (6)$$

However, when the sample is split on variable X_2 into 2 groups, one with $X_2 = 1,2$ and the other $X_2 = 3,4$, for the latter group this regression is

$$Y = a_3 + b_3X_1 + u_3 \qquad (7)$$

and if, in addition, this group is split on variable X_4, then for $X_4 = 4$ or 5, the regression of Y on X_1 is

$$Y = a_7 + b_7X_1 + u_7. \qquad (8)$$

This is illustrated in Figure 1.

If we also have

$$\Sigma u_1^2 > (\Sigma u_2^2 + \Sigma u_3^2) \qquad (9)$$

and

$$\Sigma u_2^2 > (\Sigma u_4^2 + \Sigma u_5^2) \qquad (10)$$

and

$$\Sigma u_3^2 > (\Sigma u_5^2 + \Sigma u_7^2) \qquad (11)$$

it is clear that the various effects of X_1 on Y as revealed by the differences in the slopes (b_i) and the intercepts (a_i) are associated with the joint occurrence of the conditions denoted by the indicated values of variables X_2, X_3 and X_4. Thus one must devise a means of searching various sample sub-groups in order to learn whether these differential effects exist and what forms they take, under adequate constraints to reduce the probability of detecting differences which are spurious. It is this problem which lends itself to solution via a formal sequential decision process programmed to run on a

TABLE 2.

Analysis of Variance for Regression

Source of Variation	D. F.	Sum of Squares	Mean Square
Regression estimates around $\bar{Y}$	1	$\sum_{i=1}^{N} (\hat{Y}_i - \bar{Y})^2$	$(N-1) r^2 s_y^2$
Observations around regression estimates	N-2	$\sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2$	$\frac{N-1}{N-2} (1-r^2) s_y^2$
Observations around $\bar{Y}$	N-1	$\sum_{i=1}^{N} (Y_i - \bar{Y})^2$	s_y^2

Where $r^2 = \frac{[\Sigma (X-\bar{X})\ (Y-\bar{Y})]^2}{\Sigma\ (X-\bar{X})^2 \Sigma (Y-\bar{Y})^2}$

$s_y^2 = \frac{\Sigma (Y-\bar{Y})^2}{N-1}$

$\hat{Y} = \bar{Y} + b_{yx} (X-\bar{X})$

$b_{yx} = \frac{\Sigma (X-\bar{X})\ (Y-\bar{Y})}{\Sigma (X-\bar{X})^2}$

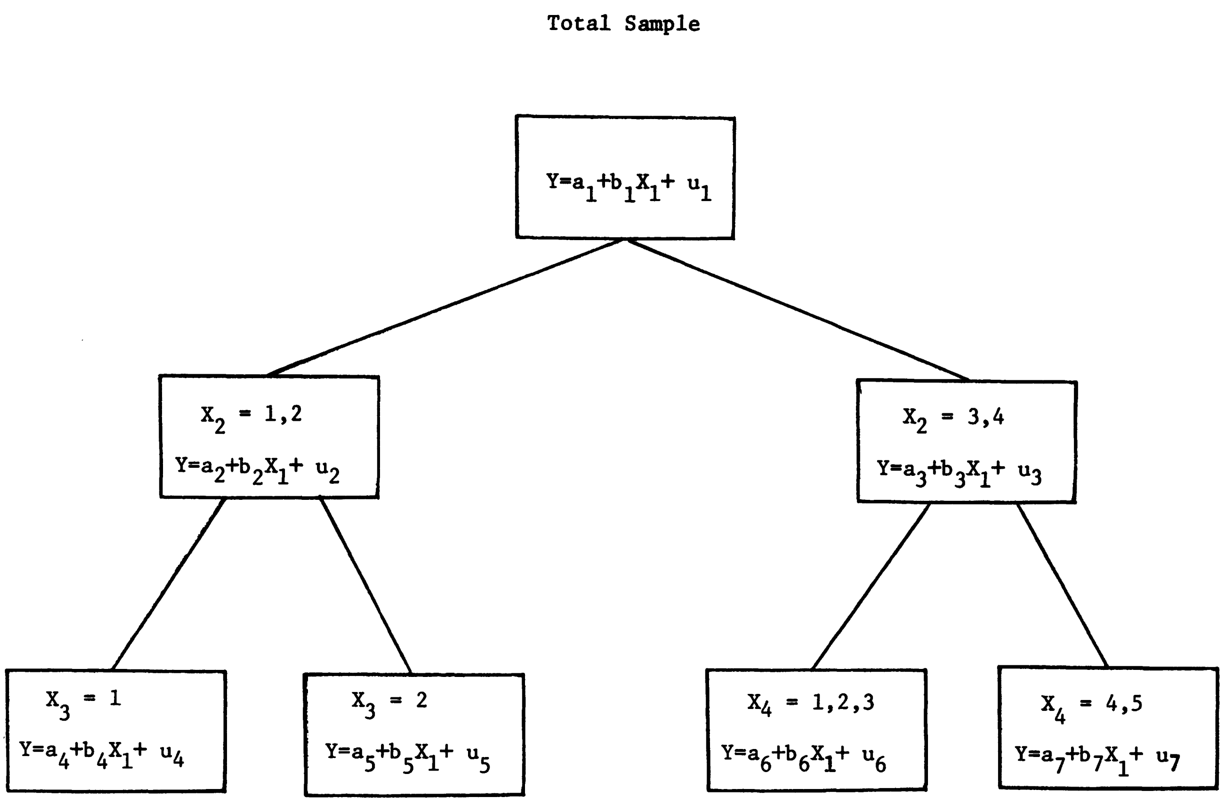


Figure 1. Differential Effects of X_1 on Y_1.

TABLE 3.

Logical Models for Two Dichotomous Predictors and a Trichotomous Dependent Variable

Configuration	$\bar{A}\bar{B}$	$A\bar{B}$	$\bar{A}B$	AB Y=0	 Y=1	 Y=2
Column(j) / Row(i)	1	2	3	4	5	6
1	0	0	0	A[a]	CS	CS
2	1	0	0	CS	BICOND	(M)BICOND
3	2	0	0	CS	(M)BICOND	BICOND
4	0	1	0	CS	A	(M)CS
5	1	1	0	A	SC	R
6	2	1	0	(M)CS	R	(M)BICOND
7	0	2	0	CS	(M)CS	A
8	1	2	0	(M)CS	A	(M)SC
9	2	2	0	A	(M)SC	SC
10	0	0	1	CS	A	(M)CS
11	1	0	1	A	SC	R
12	2	0	1	(M)CS	R	(M)BICOND
13	0	1	1	EXOR	SC	A
14	1	1	1	SC	A	CS
15	2	1	1	A	CS	BICOND
16	0	2	1	(M)EXOR	R	(M)SC
17	1	2	1	R	CS	A
18	2	2	1	(M)SC	A	SC
19	0	0	2	CS	(M)CS	A
20	1	0	2	(M)CS	A	(M)SC
21	2	0	2	A	(M)SC	SC
22	0	1	2	(M)EXOR	R	(M)SC
23	1	1	2	R	CS	A
24	2	1	2	(M)SC	A	SC
25	0	2	2	EXOR	(M)EXOR	SC
26	1	2	2	(M)EXOR	EXOR	SC
27	2	2	2	SC	SC	A

[a] A means additive; CS means cumulative upwardly and substitutive downwardly; SC is the inverse of CS; (M) means "modified;" BICOND means "biconditional;" R means reversal and EXOR refers to "exclusive or". Table adapted from Sonquist (1969a).

large scale computer.

Extending the Search Algorithm - A Multi-step Look-ahead

Experimentation with the original algorithm to determine its behavior under known conditions was carried out using contrived data. This process of working out tests under a variety of different conditions led to the development of a typology of multivariate models. From this viewpoint, multivariate interactive and additive AID and MCA models could be viewed as eighty-one variants of the same basic structure, which, in its simplest form (two dichotomous predictors and a trichotomous dependent variable) could be defined almost entirely in terms of the fundamental operators of Boolean algebra, (see Table 3)[1], Eliminating permutations and inverses reduces the number of models to seven non-trivial ones, (see Table 5)[2].

This typology also delineated exactly certain limitations of the original AID algorithm, some of which could not have been noticed earlier. For instance, while the published procedure was found to be capable of dealing adequately with many two-way interactions, others were identified as being difficult for it to deal with (Sonquist and Morgan, 1964). These were seen to consist of interactive relations characterized by consistency, i.e., by balance or symmetry. One such example is the biconditional model, illustrated in Table 5.

Table 4.

Occurrence of Effect "C" Associated with the Joint Occurrence or Joint Absence of Two Causal Factors, A and B.

Condition "A"	Condition "B" B	Not-B
A	C	Not-C
Not-A	Not-C	C

1 A more extensive discussion of these models is given in Sonquist (1969a)

2 The eighth is a constant in all cells.

Although the frequency with which variants of this model actually occur in real data is not known, it is notable that at least one realization has received considerable attention in the recent sociological literature, the concept of status inconsistency.[1] The problem of developing analysis techniques to deal with this class of models is of importance for economic, educational and psychological research as well as for sociology.

It can be seen that the earlier sequential partitioning algorithm which examines only the "zero-order" effects of A and B separately could not discover the consistency effect in these data. There are really two A "effects" and they cancel each other out in the total group. Moreover, the additive assumptions implied by the choice of Multiple Classification Analysis would also tend to conceal the real state of the world.

However, the extended AID algorithm partitions the sample tentatively, first on one causal variable and then on the other (as well as making tentative partitions on other variables). The actual partition is made so as to maximize the effects of several successive partitions. This makes it possible first to reveal a consistency effect to the analyst by means of appropriate output, then to make an appropriate partition, and finally, to continue with the rest of the sequential search procedure.

In general, such a two-split scanning algorithm appears capable of providing information adequate for the analyst to identify the two-way interactions existent in the data. It also appears able to provide leads or clues to the existence of three-way interactions. This is seen to be a simple extension of the way in which the present algorithm provides clues to the existence of two-way interactions. Thus, an algorithm which examines the cross-classifica-

1 For an example, see Blalock (1966).

TABLE 5

Seven Logical Models for Two Dichotomous Predictors and a Trichotomous Dependent Variable

Key = Cell Averages

H = high cell mean
L = low cell mean
M = middle cell mean

1. Univariate

	B_1	B_2
A_1	H	H
A_2	L	L

2. Cumulative

H	L
L	L

3. Modified Cumulative

H	M
L	L

4. Additive

H	M
M	L

5. Biconditional

H	L
L	H

6. Modified Biconditional

H	M
L	H

7. Reversal

H	M
L	M

tion of p predictors simultaneously appears able to reveal terms composed of p raw variables regardless of the symmetry of the term. However, such an algorithm also appears capable of revealing a term involving p + 1 raw variables if the term is asymmetric.

For instance, if we have the three variable asymmetric model,"if A and B and C, then Y = 0, otherwise Y = 4," the algorithm using a two-split strategy would produce the sequence of partitions illustrated in Figure 2.

Of course the amount of computing required to search out combinations of three or more variables increases as an exponential function of the number of variables considered simultaneously. Hence constraints have to be put on the process to permit the eliminations of unpromising leads and thus the examination of the subsequent partitions. However, this does not appear necessary in the three variable case.

Extending the Search Algorithm - Premium for Symmetry

The original algorithm represented a step in the direction of specifying a statistical model so it fits the data rather closely. The introduction of the look-ahead principle moves further in this direction. However, it can be anticipated that increasing the "wiggling" ability of a model being fitted will also increase the probability of the analyst basing his theoretical model on data largely reflecting idiosyncratic characteristics of the sample under investigation. Hence additional constraints are needed that would tend to guard the analyst against over-fitting his model. One such constraint is a premium for symmetry. Thus, when a look-ahead is employed a capability can be provided for increasing the probability that if a given predictor is used to make a partition in a given way on one branch of a partition sequence it will also tend to be used similarly in the parallel branch. This principle of constraint toward symmetry is illustrated in Figure 3.

Once groups four and five have been created using variable B the symmetry question arises. The proposed partition of group two, the "Not-A's" into new groups could be accomplished using, say, variables B, C, or D, but not variables E, F, or G, since the latter show insufficient explanatory power. In the previous algorithm, the choice of a variable on which to base a partition would have been to compare B, C, and D and then choose the one capable of producing the greatest reduction in the unexplained sum of squares. The present proposal would alter the algorithm to require that if C or D were chosen over B, it would have to achieve a certain ratio of explanatory power when compared with that resulting from a partition based on B identical to the one already performed in the parallel group.[1] The comparison ratio would be supplied by the analyst at the time of execution of the program. Setting the ratio to 1.0 would simply cause the regular algorithm to take effect. Setting it larger than 1.0 would bias the algorithm toward symmetry; setting it at less than 1.0 would tend to prevent symmetry. For instance, a value of 1.25 would require that a non-symmetric partition explain 25 percent more variation than a symmetric partition in order to be actually used in a split.[2]

Preliminary Examination of the New Algorithm

As a preliminary investigation into the power of the new algorithm with respect to the look-ahead and covariance options, tests were made using all combinations of the seven logical models of Table 5 applied to both means and slopes (see Tables 6a and 6b).[3]

---

1 In the case of dichotomous predictors the identical partition is the only possible one. This is not the case where the variable has three or more classes. We focus on total symmetry, i.e., forming identical sub-groups based on the same predictor.

2 Appropriate values of the symmetry premium for actual use are yet to be worked out by experimentation.

3 For each of the means model in Table 5a, a dependent variable was generated, $Y_i = a + e_i$, $i = 1,\ldots,100$, where (a) represents the mean of a given cell (25 observations were generated for each call) and e_i is a random error $e_i \sim (N(0,.5)$. The e_i remains constant for all 7 models. A covariate $x_i \sim N(0,1)$, $i = 1, \ldots,100$, was also generated, and for all 49 combinations of means and slopes the dependent variable $y_i = a + bx_i + e_i$, $i = 1, \ldots,100$, was formed, where (a) and (b) are the mean and slope values for the appropriate cell. The means-alone model, $y = a + e_i$ was also evaluated, making the total number of experiments fifty-six.

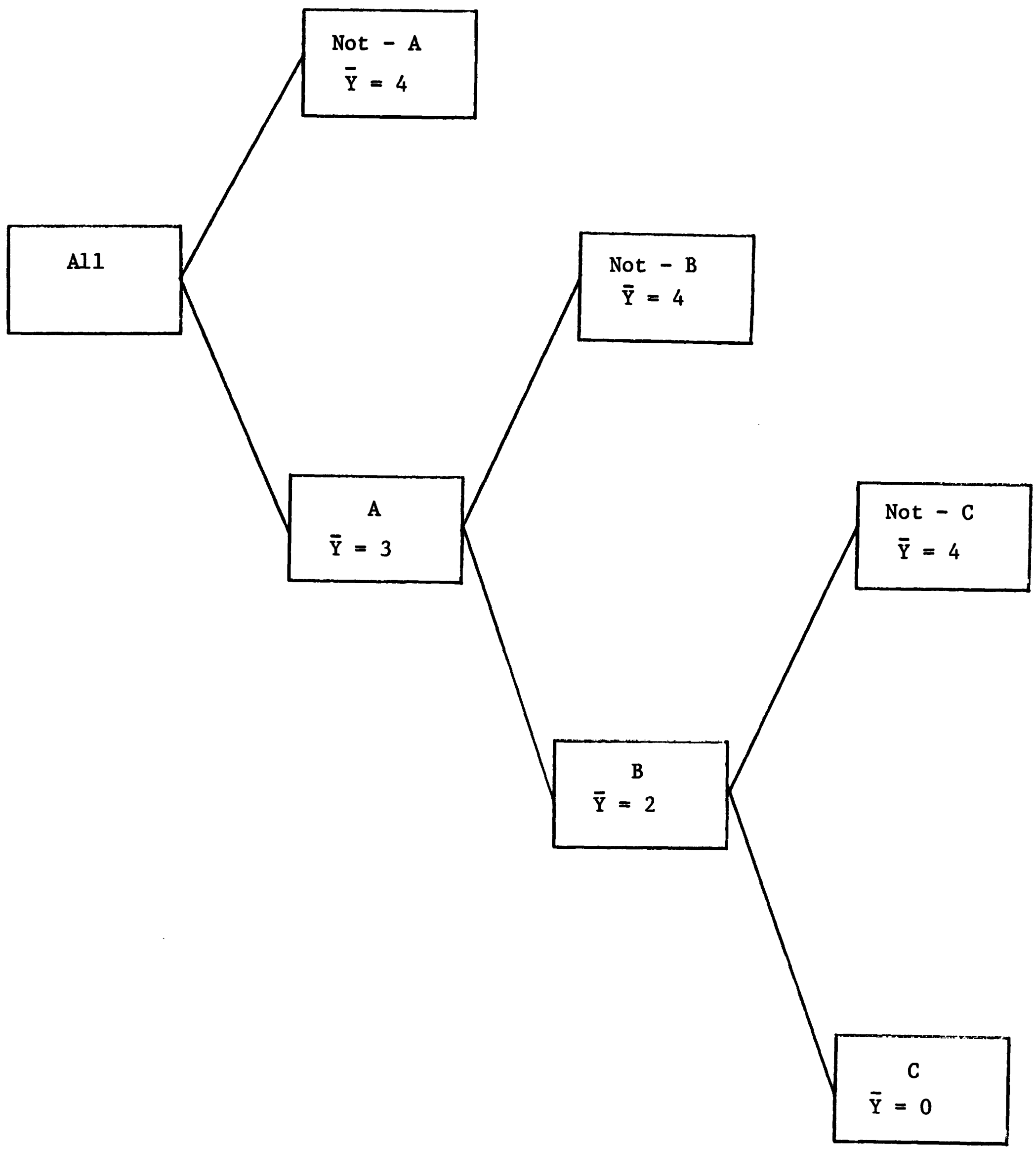


Figure 2. ABC Implies Y = 0, Else Y = 4

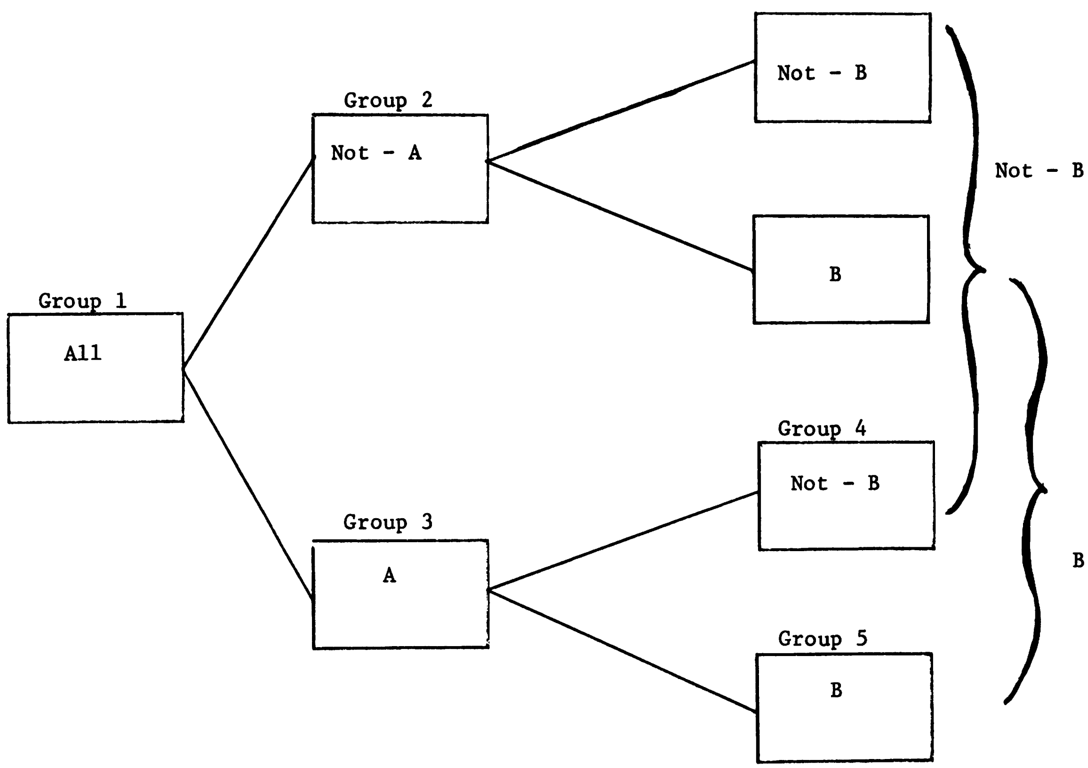


Figure 3. Symmetric Partition.

TABLE 6a

7 Logical Models Applied to Means

1. Univariate

	B_1	B_2
A_1	10	10
A_2	0	0

2. Cumulative

10	0
0	0

3. Modified Cumulative

10	5
0	0

4. Additive

10	5
5	0

5. Biconditional

10	0
0	10

6. Modified Biconditional

10	5
0	10

7. Reversal

10	5
0	5

TABLE 6b

7 Logical Models Applied to Slopes of Y on X

1. Univariate

	B_1	B_2
A_1	1	1
A_2	-1	-1

2. Cumulative

1	-1
-1	-1

3. Modified Cumulative

1	0
-1	-1

4. Additive

1	0
0	-1

5. Biconditional

1	-1
-1	1

6. Modified Biconditional

1	0
-1	1

7. Reversal

1	0
-1	0

Since this was a preliminary study, "masking" factors such as intercorrelations and noise in the predictors were not considered.[1] However, to provide a test of the algorithm's ability to pick out explanatory factors from a noisy background two uncorrelated dichotomous dummy variables were generated along with the two variables defining the model. Each experiment was made using the two real factors and the two noise factors as predictors, utilizing the "look-ahead" with two splits (creation of three groups).

For the means-alone cases, the sum of squares between the three terminal groups was maximized, i.e., from Table 1,

$$\sum_{i=1}^{3} N_i \bar{Y}_i^2 - N\bar{Y}^2 \qquad (12)$$

For the means and slopes cases, the quantity maximized between the three goups was the sum of squares resulting from both the reduction in means and regression, i.e., from Table 2,

$$\sum_{i=1}^{3}\left[N_i\bar{Y}_i^2 + (N_i-1)r_i^2 s_{yi}^2\right] - \left[N\bar{Y}^2 + (N-1)r^2 s_y^2\right] \qquad (13)$$

As was expected, with the exception of the biconditional model, the look-ahead with two splits yielded results similar to those from no look-ahead; the final groups in both cases were identical, but occasionally the look-ahead would lead to splits on the variables in reverse order from a parallel analysis using no look-ahead.

The biconditional-means model still presented some problems, however. For the no-slope version of the model as well as the univariate, cumulative, modified cumulative, additive and reverse slope models, the algorithm with no look-ahead could make incorrect splits; that is, it could use the dummy variables by mistake. it is significant, however, that the look-ahead identified the models correctly. However, it occasionally made subsequent partitions on the dummy variables. This proved to be capable of remedy by an adjustment of the reducibility criteria controlling the termination of the partitioning process.[1] Proper choice of this criterion still permits legitimate partitions to take place, but prevents subsequent spurious ones. For the no-look-ahead case, this fraction p_o apparently should be in the range $.005 \le P_o \le .008$. For the one-step look-ahead (two splits) this fraction P_1 apparently has a lower bound of $P_1 \le .016$. The look-ahead apparently does the job it was designed for.

Our findings from this initial experimentation with covariance models suggest that the differences in means may be much more powerful in determining what split is to be made than are differences in slopes. In fact, in most of the cases where (a) the two groups had no differences in means and (b) one of these groups had slope zero, the two regressions resulting from tentative partitions were not sufficient to meet the reducibility criterion with, or without, a look-ahead. For instance, in the example given below in Table 7, the algorithm would split the sample on variable A, but would not split either of the resulting groups.

Table 7.

An Example of a Remaining Problem

	Means B_1	Means B_2		Slopes B_1	Slopes B_2
A_1	10	10	A_1	1	0
A_2	0	0	A_2	0	-1

[1] Further work is to include tests with predictors of varying levels of intercorrelation and skewness as well as the assessment of the algorithm activity to deal with correlated "noise."

[1] The criterion is the fraction of the total sum of squares from the total input sample that a split (or sequence of splits in the look-ahead case) must explain in order for the split actually to be made. The maximized function [equation (12) or (13)] must be greater than p_i times TSS. The behavior of the algorithm is in keeping with previous results from the original algorithm. For $p_o <$ .005, there is a tendency to split on the dummy variables after the correct splits have been made. For $p_1 <$ .016 the look-ahead would occasionally split on a dummy variable even before making the correct splits.

While this result is not entirely unexpected, it may imply that maximizing differences in means alone or in regression slopes alone may be a more powerful tool than the combination of the two.[1] Further experimentation with these models using only the slopes is obviously indicated.

Significance of These Findings

This extended algorithm represents a continuation of our attempts to develop better methods for adequate handling of the problem inherent in applying multivariate statistical techniques in the analysis of cross-sectional survey data with large numbers of cases (1000 or more). The covariance capability is relevant for the analysis of panel as well as one-time cross-sectional data. The increasing volume of survey and other nonexperimental data and the growing needs of researchers in many disciplines to utilize it in solving practical problems makes the development of adequate data-handling methods even more imperative. The procedures are frankly oriented toward the inductive phases of research in which model-building is the objective rather than toward the deductive model-testing phase. It is felt that a focus here will be of benefit both to theory builders and to those who must find immediate solutions to pressing problems.

---

1 It is possible that this behavior results largely from the sizes of the mean and slope effects relative to the noise level. This remains a problem for further exploration.

References

Andrews, F.M., Morgan, J.N., and Sonquist, J.A. Multiple Classification Analysis. Ann Arbor: University of Michigan, Institute for Social Research. 1967.

Aptakin, Peter. Automatic Interaction Detector. New York: Service Bureau Corporation, Computing Sciences Division. 1965.

Arpi, Bo. En Modell for Optimal Marknadssegmentering. Lund Business Studies No. 4. Studentlitteratur Lund. 1967.

Biervert B., Dierkes, M., and Walzel, A. Automatischer Split zur Optimalen Kombination Erklarender Faktoren. Forschunsstelle fur Empirische Sozialökonomik und Rechenzentium, University of Cologne, March, 1968.

Blalock, H.M. The identification problem and theory building: The case of status inconsistency. Amer. Soc. Rev., 31, 1, 52-61.

Campbell, Robert H. Modifications to the Automatic Interaction Detector COMCOM/Simulation Memo No. 26. Program operational on the MIT-FMS system. Cambridge: Massachusetts Institute of Technology. 1965.

Caplan, N., Lippitt, R., Gold, M., Mattick, H., Suttles, G., and Deshaies, D. The Outcome of the Chicago Youth Development Project. New York: Boys Club of America Convention. 1966.

Carlson, W.L. Identification of the Problem Driver from Driver Records: A Preliminary Analysis. Ann Arbor: Highway Safety Research Institute, University of Michigan. 1968.

Gensemer, Bruce L., Lean, Jane A., and Neena, William B. Awareness of marginal income tax rates among high-income taxpayers. National Tax Journal, XVIII, September, 1965, 258-267.

Kay, Kevin, Automatic Interaction Detector: "AID" Translater into CDC 3600 Fortran and Compass, Technical Report 46. East Lansing: Michigan State University, Computer Institute for Social Science Research. 1966.

Land, K.C. The Sociology Program Library, A User's Manual. Austin: Population Research Center, Department of Sociology, University of Texas. 1968.

Lord, F.M. Elementary models for measuring change. In Problems in Measuring Change, C.W. Harris (ed.) Madison: University of Wisconsin Press, pp. 21-39. 1962.

Marks, E. NCHS AID Program. Unpublished manuscript. Wharton School, University of Pennsylvania. 1965.

Morgan, J.N., and Sonquist, J.A. Problems in the analysis of survey data: And a Proposal. Journ. Amer. Stat. Assoc., 1963(a), 58, 415-434.

Morgan, J.N., and Sonquist, J.A. Some results from a non-symmetrical branching process that looks for interaction effects. Proceedings of the Social Statistics Section, American Statistical Association, 1963(b).

Ross, J., and Bang, S. Predicting the adoption of family planning. Studies in Family Planning, No. 9. 1966.

Sarlvik, Bo. Socio-Economic Predictors of Voting Behavior: Research Notes from a Study of Political Behavior in Sweden. Unpublished Manuscript. 1968.

Snowbarger, Marvin. An Interaction Analysis of Consumer Durable Expenditures. Unpublished Ph. D. thesis, University of Michigan. 1966.

Sonquist, J.A. Automatic Interaction Detection and Multiple Classifiecation Analysis: The Validation of a Search Strategy. Unpublished Ph. D. dissertation, University of Chicago. 1969(a).

Sonquist, J.A. Finding variables that work. Public Opinion Quarterly, 1969(b), Vol. XXXIII, No. 1, pp. 83-96.

Sonquist, J.A. Simulating the research analyst. Social Science Information, 1967, Vol. 6, No. 4, p. 207-215.

Sonquist, J.A., and Morgan, J.N. The Detection of Interaction Effects. Ann Arbor: Survey Research Center Monograph No. 35, Institute for Social Research, University of Michigan. 1964.

Walker, H.M., and Lev, J. Statistical Inference. New York: Holt, Rinehart and Winston. 1953.

A COMPUTER LANGUAGE FOR THE ANALYSIS OF VARIANCE

David J. Armor, Harvard University

Computer processing methods in the social sciences have undergone radical change during the sixties. The change has been along two major dimensions. First, increased computing capability has led to considerable development in numerical analysis and statistics. It has become possible to handle such complicated computational problems as matrix inversion and characteristic equation solutions. The consequence is increased stress on multivariate methods which have been relatively ignored to this point. Among social scientists, interest has grown in the potential applications of such methods as multiple regression, factor analysis, and discriminant function analysis. Aside from their conceptual relevance, for the first time it is practical for him to consider such methods since "canned" programs remove the requirement of understanding the necessary mathematical operations. Several programs for the multivariate analysis of variance, one of the more complicated methods, have become widespread.[1]

Second, at the beginning of this decade the few social scientists who analysed their numerical data by computer relied on fairly simple "canned" programs, such as those supplied in Dixon's BMD series.[2] While many of these programs provided the social scientist with considerable computing power, they were very often quite restricted in their assumptions about data format, transformations, the number of parameters, and the like. As more social scientists began utilizing the computer, the demands for flexibility and generality spurred efforts to develop more comprehensive computer systems for data analysis. Some of these systems have taken on the character of "natural" languages which make it possible for the researcher to request many different and complex analysis with a minimum of knowledge about computer technology.[3]

I would like to describe the development of a data analysis system called Data-Text which takes advantage of these two trends.[4] In particular, I would like to describe the way in which the Data-Text is a natural computer language with instructions expressed in words or terms which can be understood by the average social science researcher. The language allows for a wide variety of data input and transformation. Moreover, the statistical analysis routines, particularly those for the analysis of variance, contain many options and combinations which are ordinarily found only in separate canned programs. The existence of these features makes it relatively easy for social scientists to request quite complicated analyses of variance on a great variety of data types.

Before presenting the analysis of variance language, it will be necessary to describe briefly the overall Data-Text language. All of the statistical routines in Data-Text take advantage of the preliminary Data-Text features for the definition and labeling of variables. Rather than describe the whole language in detail, an example will be used to illustrate the relevant features. Let us assume a study involving any number of subjects and several sets of variables: background variables, such as sex, age, education and an ability test score; treatment variables, such as drug and stress conditions; and dependent variables, such as blood pressure, pulse rate, and a test battery of 10 yes-no questions measuring anxiety symptoms.

We assume that the data is punched with two cards per subject (any number would be possible). If UNIT refers to a subject identification field, and if COL refers to column number (and COL.../n)

```
*DECK      SAMPLE DATA FOR ANOVA
*CARD(1)/  UNIT = COL(1-4), CARD = COL(5)
*CARD(2)/  UNIT = COL(2-5), CARD = COL(79)
*SEX = COL(9/1) = SEX OF RESPONDENT(MALE/FEMALE)
*EDUC = RECODE(A)  COL(36-37/2) = EDUCATION(GRAM/HS/COLL)
*AGE = COL(10-11/1)
*CODE(A) = (1-4=1/5-7,9=2/8,10-12=3/OTHERS=BLANK)
*DRUG = COL(15/2)=DRUG CONDITION(PLACEBO/ASPIRIN/CODEINE)
*BLP(1-3) = COL(41-43,45-47,49-51/2) = BLOODPRESSURES
*VAR(4) = COL(6/1)+1 = STRESS(LOW/HIGH)
*VAR(6) = COL(71-72/1)=ABILITY TEST SCORE
*VAR(7) = (VAR(6)/AGE) *100 = IQ
*PULSE(1-3) =COL (62 63,64-65,66-67/2) IF DRUG = 2, 3 = PULSE RATES
*ITEMS(1-10) = COL( 51(10)/1) = ANXIETY ITEMS
*ANXIETY = SUM ITEMS(1-10) = ANXIETY INDEX
*PRINT UNIT IF VAR(7) GREATER THAN 200
*COMPUTE FREQUENCIES
*COMPUTE CORRELATIONS(6,7,ANXIETY BY BLP(1-5), PULSE(1-3)), TEST
*COMPUTE FACTORS (ITEMS(1-10)), MAX = 3, NOSCORES
*COMPUTE REGRESSION (4, ANXIETY ON SEX, EDUC, 6, 7,), RESIDUALS
*COMPUTE CROSSTABS (4, DRUG BY  EDUC BY SEX), TEST, GAMMA
*COMPUTE PLOTS (7,  ANXIETY BY BLP(1-3))
(data Deck or a READ TAPE instruction)
*END
```

refers to columns in the nth card), then the following Data-Text instructions are sufficient to define the variables, with the transformations indicated, and to compute a variety of statistical analyses (each line might represent either a punched card or a line typed at a console):

Each of the statistical analyses requested by the *COMPUTE instructions is a separate routine loaded by the Data-Text system. This makes it possible to perform several quite distinct analyses with a single computer run. On a console, it would be possible to get the results of each analysis before requesting another. The results are printed out making full use of the variable numbers, variable names, and category names.

For example, one face of the three-way crosstab appears as:

One of the keys to the simplicity of the *COMPUTE instructions is the choice of default conditions which correspond to the most common usage of a given analysis. For example, in all routines missing observations are always assumed by default; in regression, stepwise is assumed; in factor analysis, principle components is assumed. The options available are specified separately by users who desire them. Aside from its simplicity, this approach avoids cluttering up a print-out with information not useful or not meaningful to a user (e.g., a full inverted correlation matrix of the independent variables in regression analysis).

The *COMPUTE instructions for the analysis of variance routine in Data-Text are somewhat more complicated than the instructions presented so far. The main reason is that the routine is fairly generalized so that a wide variety of designs can be handled. The routine is designed to handle a large number of factors (or classificatory variables) with any number of levels on each, the main restriction being the available computer memory.

CONTINGENCY TABLE 1

CELL PERCENTS BASED ON COLUMN SUMS
SUBTABLE OF UNITS WITH MALE ON SEX OF RESPONDENT

		EDUCATION GRAM	HS	COLL	TOTAL	PERCENT
VAR(4) STRESS	LOW	22.2 4	48.6 18	68.2 15	37	48.1
	HIGH	77.8 14	51.4 19	31.8 7	40	51.9
TOTAL		18	37	22	77	
PERCENT		23.4	48.1	28.6		100.0

CHISQUARE = 8.388 WITH 2 DF (SIGNIFICANT AT THE .015 LEVEL)
GAMMA = -.529

The routine handles factors which are crossed or nested, or any combination of such. In these designs, UNIT's (or subjects) are assumed to be nested within the cells generated by the classification structure. For example, using the variables derived in our example,

*COMPUTE ANOVA (SEX BY EDUC), ANXIETY

takes the anxiety score as the dependent variable and sex and education as the factors. The operator "BY" indicates a factorial design, and the number of levels is taken from the number of categories given in the variable definition of each. Thus, this would be a 2 X 3 factorial design with replications.

If several univariate analyses with the same design are desired, the instruction would be

```
*COMPUTE ANOVA(SEX BY EDUC),VAR(4,7),ANXIETY
```

Three separate anovas would be carried out, one at a time, for stress, IQ, and anxiety.

There are two general default assumptions in these examples which apply to certain other examples as well. As we said, UNIT's (subjects) are assumed to be nested within each cell. There may be, however, unequal numbers of observations per cell, so that non-orthogonal factorial designs can be analyzed. The method used is that of unweighted means.[5]

The second default is fixed effects; i.e., the levels of each factor are assumed to be a universe of factor levels. If the levels are sampled, the sampling fraction -- or the option R (random) if sampling from an infinite universe -- can be placed after the VAR number:

```
*COMPUTE ANOVA(SEX BY EDUC(R)),ANXIETY
```

would cause education to be treated as a random effect, and this analysis would be handled as a mixed model.

Factors which are nested are indicated by the operator "within". Assume we defined two additional variables as follows:

```
*STATE=COL(77/1)=(NY/MASS)
*CITY=COL(78/1)=(ALBANY/NYC/BOSTON/SPRING)
```

City is nested within state, and the ANOVA request

```
*COMPUTE ANOVA(CITY WITHIN STATE),ANXIETY
```

would cause the appropriate nested design analysis. In these designs UNIT's are assumed to be nested within cells.

Any number of factors (up to 10), either fixed or random, can be combined in an expression and the correct analysis will be carried out. Parentheses are used for clarity of the nesting relationships:

```
*COMPUTE ANOVA(EDUC BY(CITY WITHIN STATE)),ANXIETY
```

would be a three-factor design, with education crossed by both city and state.

The testing of effects is made possible by implementing the Tukey-Cornfield rules for finding the correct denominators for F-tests.[6] These rules cover most combinations of fixed or random and crossed or nested factors. Special options are available in the event that a denominator cannot be found for a given test.

In many behavioral science applications, subjects are measured several times, with a subject becoming his own control. Examples are survey panel studies and learning experiments. These set-ups are often termed "repeated measure" designs. They present special problems for the Data-Test system. For example, assume that the pulse rate variables, PULSE(1-3), were actually measured at three different times, and it is desired to test for changes over time. The usual approach is to assume a factorial design with UNIT's crossed by time and with one observation per cell. However, the time factor is not an explicit Data-Text variable, as were the factors in our previous examples; the levels of time are implicit in the existence of three pulse rate measures. Moreover, the dependent variable, pulse, is not a single Data-Text variable. To indicate this type of design, we provide special *FACTOR and *MEASURE definition instructions.

```
*COMPUTE ANOVA(UNIT BY A),MEASURE(1)
*FACTOR(A)=TIME(TIME1/TIME2/TIME3)
*MEASURE(1)=PULSE(1=A1/2=A2/3=A3)=PULSE RATES
```

The factor instruction gives the structure and labels for the factor, and the measure instruction relates the dependent variable to the levels of the factor (A1,A2,A3). In these designs, the default is that UNIT's are a random effect, so that this example represents a mixed model.

If we assume that there was a second repeated measure factor, say a treatment condition of some kind, and additional pulse rate measurements, then the following instructions would specify a three factor design with UNIT's crossed by time and by treatment:

```
*COMPUTE ANOVA(A BY B),REPEATED MEASURE(1)
*FACTOR(A)=TIME(TIME1/TIME2/TIME3)
*FACTOR(B)=TREATMENT(COND1/COND2)
*MEASURE(1)=Pulse(1=A1,B1/2=A2,B1/3=A3,B1/     *
*   4=A1,B2/5=A2,B2/6=A3,B2)=PULSE RATES
```

The option REPEATED on the *COMPUTE instruction has the effect of crossing UNIT by every factor within the parenthetical design specification.

More complex designs can be requested which have some factors crossed by UNIT and other factors within which UNIT's are nested. An example might be

```
*COMPUTE ANOVA((UNIT WITHIN SEX)BY A),    *
*                   MEASURE(1)
*FACTOR(A)=TIME, etc.
```

In this case we have the repeated measure assessment -- UNIT's by time -- carried out on both males and females. Thus, UNIT's are nested within sex, but time is crossed by sex and by UNIT.

All of the designs discussed can have covariates specified, and the appropriate analysis of covariance will be computed.

```
*COMPUTE ANOVA(SEX BY EDUC),ANXIETY/VAR(6)
```

will treat VAR(6), ability score, as a covariate. The results include the regular tests for anxiety, the covariance or regression test, and tests for the anxiety effects after adjusting for VAR(6). A covariate (or any number of covariates) can be specified on any of the designs discussed earlier.

The routine can also handle the generalized multivariate case. If one has several dependent variables which are to be tested simultaneously, (the 10 anxiety items, for example), then the option MANOVA on the following instruction will cause a multivariate analysis:

```
*COMPUTE MANOVA(SEX BY EDUC),ITEMS(1-10)
```

The results include a multivariate test for each effect implied in the design using the likelihood ratio criterion.[7] The univariate tests are also given. The MANOVA option can be used with any of the designs discussed so far, including the covariance case.

The output display in both the covariance and the MANOVA cases include the appropriate vectors of cell and marginal means, and the within-cell standard deviations and correlation matrices. The instruction

```
*COMPUTE CROSSTAT(SEX BY EDUC),ITEMS(1-10)
```

will produce just this display part without the univariate and multivariate testing. The only difference is that all marginal means will be weighted if cell N's are not equal. This option makes it easy to get basic statistics and correlation matrices within a complex grouping structure.

As in the other statistical routines in Data-Text, considerable attention is given to the problems of missing observations in the various analysis of variance designs. Missing observations on a single dependent variable are handled by treating the design as non-orthogonal; i.e., unequal numbers of cases per cell. If the problem is multivariate, as in MANOVA, CROSSTAT, or covariance, missing observation cross-products matrices are accumulated within cells and pooled to form an estimate of the population correlation matrix.

For the repeated measures case, the problem is somewhat more complicated since a missing observation is tantamount to a missing cell. The default procedure adopted is one of iterative least squares estimation of missing values for a given UNIT using the marginal means for that UNIT.[8] The user may select an option to omit UNIT's with missing observations for both the repeated measure and the multivariate cases.

There is not sufficient space to show a detailed example of the planned printed output of the results; instead, I shall summarize the major contents of the output for the various designs.

1) For non-repeated measure designs, the standard deviations, and counts will be displayed in a tabular form similar to the CROSSTAB table shown earlier. Full use will be made of variable and category labels. All possible marginal means will also be displayed.

2) If requested, effect estimates will be shown in tabular form similar to the display of means.

3) In the MANOVA, CROSSTAT, and covariance cases, the within-cell correlation matrix of the dependent variables and covariates will be printed. In the covariance case, adjusted effect estimates will also be displayed.

4) An analysis of variance table will be produced showing source, degrees of freedom, sums of squares, variance components, F-ratios (where possible), and significance levels. In the covariance case, both the original and adjusted anova tables will be shown. In the MANOVA case, the likelihood ratio criterion will be printed.

5) If an F-test cannot be found for a given effect using the Tukey-Cornfield rules, a table of expected mean squares will be produced to aid the researcher in making pseudo-F ratios.

Our present plans do not call for handling more complex designs. For example, the routine will not provide a solution for factorial designs with missing cells or nested designs with unequal numbers of nests. Moreover, there is no provision handling such special designs as Latin squares. Future plans do call for the addition of an option for pooling mean squares for the purpose of combining or deleting various effects in the model, and options for testing special comparisons among main effects.

Obviously, the goal of a simplified language means some sacrifice in the scope of the routine, although the present routine will handle the most common designs. The main purpose of the Data-Text system is to make complex methods available to the average social science researcher. Many researchers avoid analysis of variance because of the complexities involved in learning the computer procedures or because they must learn complex statistical terminology. Hopefully, the gain should be increased utilization of an extremely powerful technique.

Since the Data-Text project and the analysis of variance routine are still in the developmental stages, we welcome critical comments and suggestions. The analysis of variance procedures adopted are sufficiently complicated to deserve continued scrutiny and revisions when necessary.

NOTES

* The Data-Text system was developed originally under the direction of Dr. A.S. Couch. Principle associates were David Peizer and Mary Hyde. Peizer also designed the original plans for the analysis of variance routine. Principle programmers for the routine have been Rod Montgomery, Frank Benford, and Karl Deirup. Donald Rubin has given further statistical assistance with the help of staff members in the Department of Statistics, Harvard University. The current version runs on the IBM 7090/94. A project to revise the current version for the IBM 360 and other computers is being supported by an NIMH grant (MH-15884-01), with the author as principle investigator.

1. For example, the MANOVA program from Dean J. Clyde, et al, "Multivariate Statistical Programs," Biometric Laboratory, University of Miami, 1966; also Jeremy D. Finn, "Univariate and Multivariate Analysis of Variance and Covariance," Statistical Laboratory, Department of Education, University of Chicago, 1966.

2. W.J. Dixon, ed., "BMD -- Biomedical Computer Programs," Berkeley: University of California Press, 1967.

3. Examples are Jeffrey W. Bean, et al, "The Beast," Washington, D.C., The Brookings Institution, 1968; Norman Nie, "Statistical Package for the Social Sciences," NORC, University of Chicago, 1968.

4. A.S. Couch, "The Data-Text System," Departmental of Social Relations, Harvard University, 1967.

5. Henry Scheffe, The Analysis of Variance, New York: John Wiley, 1959, p. 362-363.

6. A description can be found in B.J. Winer, Statistical Principles in Experimental Design, New York: McGraw-Hill, 1962, 195-199.

7. See T.W. Anderson, An Introduction to Multivariate Statistical Analysis, New York: John Wiley, 1958, Chapter 8. The criterion is also known as "Wilks lambda."

8. George W. Snedecor and William G. Cochran, Statistical Methods, sixth ed., Ames, Iowa: Iowa State University Press, 1967, p. 317-321.

Discussion

George Sadowsky, The Brookings Institution

I would like to comment on these three papers as a user of computers and as a tool builder rather than as a statistician.

The central concern of the papers is with the problem of identifying and/or constructing homogeneous population subgroups. Armor proposes a tool for computing the analysis of variance for models of a "bread and butter" type; a successful tool of this kind would lower the cost to researchers for doing these calculations. Sonquist is concerned with explaining variance by dividing the population into subgroups -- with the implication of later forming an explanatory model so that the subgroups formed by the procedure become increasingly homogenous as the procedure is extended. Hall is concerned more with simply recognizing clusters that exist, regardless of the homogeneity of the characteristics of observations within a cluster.

Armor's paper contains an error of modesty that is worth correcting. In his introduction he states that it was the demand for high-level languages for social science computing that has caused them to evolve. My observations indicate that this is not at all the case; rather, the supply of such languages has led the demand for them, and it is to people like Armor that we owe our thanks for producing these useful tools -- especially considering the lack of academic recognition given to this activity.

Armor's paper appears to be a nice integration of a widely-used class of analysis of variance design into the existing Data-Text language. There are some syntax alternatives that I would prefer to those in his paper, but the differences are relatively minor and I agree with the spirit of his construction. A larger issue is the condition under which Data-Text type language development is itself worthwhile. The principal alternatives to such development are (1) the existence of general programming languages such as Fortran, (2) languages of high-level operators such as APL and the Lincoln Labs RECKONER, and (3) interrogative systems rather than declarative languages.

The case for the development of specialized computing tools appears to be a strong one. Current price trends within the computing industry indicate that the roles and costs of hardware and software are rapidly reversing. In this sense Data-Text can be regarded as an investment in capital which once created, has a low marginal cost of distribution (unlike hardware) and high marginal benefit. The case for declarative languages is weaker, but it is certainly true that Data-Text type tools operating in batch environments are currently very useful and will be for some time. Furthermore, the competition between the very different approaches of interactive systems and declarative languages is likely to be strong and will yield benefits to social science computing as a whole.

Other relevant points of Armor's paper are: (1) missing data is handled; having to deal with missing data is not at all aesthetically pleasing for the systems designer, but crucial to widespread and general use of the system; (2) the problem of an "escape" into more powerful languages to handle extensions is not visible (such as extensions to a Latin square design), and although manufacturing "escapes" is difficult, it is useful to try to provide open-endedness; (3) intuitive language structure such as is employed by Data-Text may be easy to use, but consideration should also be given in formulating syntax rigorously, both for definition and for potential implementation using meta-compilers.

Sonquist's paper explores several additions to his already well-known AID algorithm. I have used AID, and would place substantial value on deepening the search strategy. Although I have not -- at least knowingly -- had occasion to be concerned with a covariate in our analysis, I would think that it might be a quite useful addition.

Sonquist's paper fascinates me most because of the possibility of extending the AID algorithm even further within an interactive environment. Admitting that the detection of interaction effects is important in data exploration activities, why is it that such a process must be automatic? The primary reason for the automation of the process is that most computer centers -- both formerly and now -- operate primarily in the batch mode. Some of the implications of this environment for interaction detection are less than satisfactory for social scientists. For example, if two variables have nearly the same explanatory power for a given split, the one with the greatest explanatory power will always be chosen even though the analyst may have good reason to choose one of them *a priori* from a knowledge of the data. In the same spirit, having to set a filter parameter value to distinguish "signal" from "noise" is difficult *a priori* when one does not know the "signal to noise" ratio of his data. Observing the branching process interactively would allow the analyst to terminate his tree structure along various branches when it became apparent that no substantive explanatory gain could be obtained by going further. Furthermore, while the concept of expanding both the capacity of each node and the search strategy of the algorithm is laudatory, the cost increases exponentially. With some human guidance and pruning of the tree, the increase in cost of such improvement might only be moderate, and the benefits would be substantial.

I would suggest as an alternative a ''guided interaction detection'' algorithm embedded in an interactive environment. Such an algorithm could have several modes, such as automatic, semi-automatic, and manual. The automatic mode has already been implemented. The semi-automatic mode would be identical to the automatic mode except when the explanatory power of the best variable for any split was not decisively greater than all other explanatory variables and when the best reduction in variance obtained for any split decreased below a certain level. In the manual mode, the program would display the partitioning choices and the corresponding variance statistics for the analyst at each node and allow him to select the partition or form new candidate variables. The manual mode of operation would also allow ''backtracking'' upon the discovery of any evidence suggesting that a previous split might have been less than optimal, and it would also allow transition to any other procedure at any node or leaf of the tree -- such as covariance analysis, multiple regression analysis, or multiple classification analysis. Transfer between modes of operation could be simple and could be effected at the analyst's discretion. A ''guided interaction detection'' program would seem to be an extremely powerful tool for data analysis in the social sciences.

My first reaction to Hall's paper was to marvel at its convenience and elegance. I think it would be a very stimulating afternoon for me if I were in Palo Alto and had in my possession a set of data for analysis. Hall's approach seems to be more agnostic than AID, since it imposes no a priori structure on the data. Its main purpose appears to be to obtain cohesive or compact groups.

However, once PROMENADE or ISODATA has been used by the researcher, he may feel that only part of his analysis has been performed. While it is true that application of either PROMENADE or ISODATA causes subgroups to be formed and group profiles and other summary measures to be constructed, how then can the analyst move toward a model of the world he observes? Each group must be characterized, i.e., one might feel compelled to build models to explain group membership. Furthermore, if clustering is not of a very definite pattern -- either visually or actually -- the idiosyncracies of the data may cause group membership to be misleading. At the present time, the cost of a dedicated CDC 3100 computer with display is relatively high for this type of application. Although such costs will certainly decrease, they will not decrease as fast as those for simpler input-output devices that would support a guided interaction program. On the other hand, the experience gained by the use of any interactive system has a subjective quality which is hard for a reviewer to measure without having actually used the system. In the case of PROMENADE, the subjective impact becomes even harder to measure due to the extensive and elegant graphic capabilities of the system.

After reading both Sonquist's and Hall's papers, I found myself as a user wishing for some sort of a middle ground between them. They differ in that Sonquist's algorithm maintains an increasingly homogeneous characterization of subgroups formed, whereas Hall and his colleagues are more interested in data classification using metric criteria rather than in the characteristics of the subgroups. Data populations fall into a spectrum with regard to homogeniety. At the one end, there is highly clustered data, and at the other end, very diffuse data. I would prefer to use Hall's method on the former and Sonquist's on the latter.

Hall's interactive environment could be extended to allow the homogenizing of clusters based upon more analysis. For example, a researcher might prefer to use PROMENADE to isolate clusters of observations and then interactively analyze the characteristics of the clusters. He might also like to ask questions such as what the increase in unexplained variance might be if observations having certain characteristics in some clusters would be moved to another cluster. The answer to this question might determine whether he would choose to have the system move those observations and then make itself available for more such analysis.

The Monte Carlo experiments cited by Sonquist seem very worthwhile in evaluating the utility and power of these methods. One could conceive of a series of such experiments on data having different characteristics of diffusion -- perhaps using the same sample population. This would help to determine the sensitivity of these algorithms and their various modifications to ''noise'' in the data and it would help to determine the useful ranges of application of these systems.

VII

MEASURES AND MODELS IN EDUCATION

Chairman, MORRIS B. ULLMAN, U. S. Office of Education

FINANCIAL AND NON-FINANCIAL FACTORS AFFECTING POST-HIGH SCHOOL PLANS AND EVENTUATIONS, 1939-1965*

Walter Adams, Bureau of Applied Social Research
Columbia University

Part I. Introduction

If one were to characterize American education over the past century, representing the span for which we have reasonably adequate statistics, one might well speak of constancy accompanying change. The constant factor would be consecutive escalation in educational attainment for both sexes, and for whites and non-whites alike. The other face of this same coin, presenting itself as change, would be successive upward shifts in the modal level of educational attainment--once again for all four sex-race groups. These patterns are clearly revealed in an age-cohort analysis of educational attainment conducted recently at the Bureau of Applied Social Research. The analysis, deriving from Census data, covered the span of years from about 1880 through the early 1960s.[1]

In addition to escalation in attainment, the cohort analysis yielded a second major constant. Racial minority youth, as well as considerable proportions of majority youth, have consistently lagged a generation or two behind in the overall rise in attainment. Such relative deprivation has affected most of the one in eight of all youth belonging to racial minorities. It has affected far smaller proportions of majority youth. Nevertheless, given the relatively large size of the white majority, at any point in time the numerical total of the educationally deprived among whites has totalled three to four times the number within the non-white minority.

If escalation in educational attainment represents a major positive achievement, its enduring correlate, relative deprivation for specific population groups, represents an abiding problem. Public concern, and the concern of educators and officials as well, has consistently focused upon students who fail to reach the modal white attainment level pertaining at a particular time. Any apparent lessening of the attainment gap has been greeted with enthusiasm. In 1893, for example, we find James Blodgett, special agent for the educational statistics of the 1890 Census, exulting over the fact that, and I quote him: "The Negro race leaped from the illiterate slavery of 30 years ago, and in that time it has taken rank with the white in eagerness to go to school." In Blodgett's day the modal white level of attainment was the seventh to eighth grades. He further noted: "There was never a time when in the whole breadth of the land there was more interest on the subject of education than now. The gains in recent years have been very great. . . ." But at this point Blodgett interjected a word of caution: ". . . there is widespread complaint that the advancement of pupils is unsatisfactory."[2] Today, with a vastly altered and augmented educational establishment, the American people voice the identical mix of enthusiasm and chagrin.

Now, in the late 1960s, the modal level of attainment appears about to shift upwards once again, from high school graduation to college attendance. If we turn to Office of Education school retention rate data, we note that in the latter 1960s first-year college students represented 40% of the total of fifth graders eight years previously. The fifth graders approximated the total age-cohort. Moreover, the retention data reveal an uninterrupted rise since the immediate post-World War II years in the size of the first-year college population relative to that of high school graduates the previous year.[3] By the late 1960s, college freshmen were appreciably more numerous than the previous year's high school graduates who failed to enter college, representing a departure from the long-term trend from the late 19th century to the depression 1930s for about nine in twenty high school graduates to enter college--somewhat more of the white boys and somewhat fewer of the other three sex-race groups.[4] (Parenthetically, there is considerable evidence that Negro youth have not shared in this recent rise in the high school to college transfer rate, and there is further evidence that the principal obstacle for Negroes increasingly relates to securing the high school diploma rather than in reaching the twelfth grade.)[5]

*The research upon which this paper is in large part based was supported by the U. S. Office of Education, and was conducted under the direction of A. J. Jaffe at the Bureau of Applied Social Research.

[1]For extended discussion and detailed tabulations of this age-cohort analysis see: Jaffe, A. J. and Adams, Walter, American Higher Education in Transition, Bureau of Applied Social Research, Columbia University, New York, 1969, mimeo. See especially Sections II and III and Part I of Appendix C (Detailed Statistical Tables), plus discussion of historical educational data in Appendix D (Methodology).

[2]Blodgett, James H., Report on Education in the United States at the Eleventh Census: 1890, Census Office, Department of the Interior, Government Printing Office, Washington, D. C., 1893, pp. 21 and 23.

[3]Statistical Abstract of the United States, 1968, Bureau of the Census, U. S. Department of Commerce, Government Printing Office, Washington, D. C., 1968, Table 181, p. 126.

[4]Jaffe and Adams, op. cit., footnote 1. See also: Jaffe, A. J. and Adams, Walter, "Trends in College Enrollment," College Board Review, Winter, 1964-65, No. 22, pp. 27-32.

[5]Jaffe and Adams, op. cit., footnote 1, Section III, Parts I and II.

In general, then, we characterize college attendance, at least for a year or two, as the most significant educational "rite of passage" in the late 1960s. Our concern in the balance of this paper will be with determinants of college planning and attendance which appear to be losing a considerable portion of their earlier relevance, and with identification of residual determinants which would appear to represent the principal obstacles to college entrance today and tomorrow.

Before reporting the findings, we shall briefly identify our major data sources for factors associated with post-high school plans from the late 1930s to the mid-1960s.

Part II. The Data

Our information on college planning in the late 1930s and mid-1950s derives from a number of national sample surveys of high school students which we obtained from the Roper Public Opinion Center at Williams College. More recent data were obtained from two very similar Census Bureau surveys, conducted in Fall 1959 and 1965 as a portion of Current Population Surveys in those years. These various surveys permitted us to determine the changing relevance of socio-economic factors to college planning. The two Census surveys were followed-up in Fall 1960 and February 1967 to determine actual post-high school behavior, and to an extent we may relate socio-economic and academic characteristics to this behavior. Such relationships are somewhat tentative, however, since only immediate entrants to college were identified, and over a third of college entrants delay entrance more than a year following high school graduation. Accordingly we concentrate on the plans of the high school seniors, simply noting in passing that the limited findings for behavior appear to support the findings for plans. As far as we can determine, about eighteen in twenty college-planners eventually reach college, plus about three in twenty non-planners.[6] Plans as seniors represent a generally realistic indicator of future eventuations.

Our information on less researched and less understood, but nevertheless significant, factors associated with post-high school plans derive from a secondary analysis of the twelfth grade computer tapes from the Office of Education's 1965 benchmark Equality of Educational Opportunity survey, commonly referred to as the Coleman study. For the sample of over 90,000 high school seniors we were able to examine over forty variables for possible relationships to post-high school plans.[7]

With these few words on the data let us turn to the findings. Our initial concern is with "classic" socio-economic determinants of post-high school plans. For the whole span of years, 1939 through the mid-1960s, we have information on the occupation of the head of the senior's household, and for the 1959 and 1965 Census surveys data on family income are available in addition. The simple question is whether, with the passage of time, democratization in college-planning by socio-economic class took place--and if so, precisely when this occurred.

Part III. Socio-Economic Class and Post-High School Plans[8]

Between 1939 and 1959 it appears that the proportion of high school seniors planning on college rose modestly, or about seven per cent, from roughly four in ten to just under five in ten of all seniors. In the far briefer six-year period, 1959 through 1965, the rise was about thirteen per cent, or nearly double that for the previous twenty years. In 1965 sixty per cent of the Census Bureau sample of seniors planned on college. By February 1967 forty-seven per cent of the seniors had entered college, and we estimate that fully sixty per cent will have entered when all of the delayed entrants are accounted for. As we have noted, a few planners will fail to enter and a few non-planners will in fact enter.

We may supplement these findings with several others which appear to explain them to an extent. All socio-economic groups of seniors, as measured by the father's occupation, shared about equally in the modest rise in expectations between 1939 and 1959. At both dates very nearly twice the proportion of white-collar seniors as of blue-collar ones planned on college. Between 1959 and 1965, however, though white-collar children increased college planning by eight per cent, manual and service children increased planning nearly twice as much, or by about fifteen per cent. In a brief six years the twenty-nine per cent planning gap between the two occupational groups was reduced by a quarter.

For the 1959-1965 span family income data are available, and for this variable democratization in college planning is even more pronounced. The findings pertain whether we do or do not reconstitute the income categories to reflect distribution changes. College planning for poor seniors, those

[6] This estimate is based on the following facts: 1) About 47 per cent of Fall 1965 high school seniors in the Census Bureau survey entered college immediately; 2) A 1966 Census Bureau study of the college population indicated that over a third of first-time college students delayed entrance over a year following high school graduation. At least 60 per cent of the 1965 seniors should enter college eventually. We simply projected the planning-non-planning distribution for the immediate entrants to include the delayed entrants.

[7] For methodological details of this analysis see: Jaffe and Adams, op. cit., footnote 1, Appendix D (Methodology).

[8] For detailed analysis and tabulations relevant to this section of the present paper, see: Jaffe and Adams, op. cit., footnote 1, Sections VI and VII and Part IV of Appendix C (Detailed Statistical Tables). See also: Forthcoming volume by Joseph Froomkin, Aspirations, Enrollments, and Resources, Part I, Section 2, U. S. Office of Education, Washington, D. C., 1969.

from families with 1965 incomes of under $4,000, increased by twenty-three per cent, whereas the increase for seniors from families with incomes of $8,500 or more in 1965 was only six per cent. The planning discrepancy between the two income groups was reduced by roughly forty per cent in six years.

In general, it would seem that it is the financial component of socio-economic status, rather than the social and psychological elements, that has been most clearly losing its earlier strong relationship to college planning. In addition, since the Coleman study clearly established the strong positive relationship between above average tested academic ability or achievement and high socio-economic status, we infer that the chief residual liabilities of low socio-economic status--liabilities, that is, in terms of post-high school aspirations and eventuations --are social and psychological ones, plus academic ones.

Supplementary findings from the 1965 Census Bureau survey support our belief that the financial deterrent to college entrance has lost much of its earlier significance. When seniors not planning on college were asked to cite the chief reason for failure to so aspire, only about one in ten of white and non-white seniors alike cited lack of money or the necessity of going to work to supplement family income. Most cited alternatives to college such as learning a trade, or taking a job from choice rather than from necessity. Between one and two in ten of all non-planners cited simple lack of interest in college. Only sixteen per cent with family incomes of under $5,000 per year cited lack of funds or the financial necessity of going to work.

In brief, by the mid-1960s it would seem that the financial factor, if we can credit the seniors' testimony, was a relatively minor determinant of post-high school aspirations. The question becomes one of accounting for this finding.

Part IV. Higher Educational Innovations in Recent Years

Since the latter 1950s a number of higher educational innovations in combination would seem to explain the liberalization in access to college for lower socio-economic high school graduates. All of the innovations represent direct or indirect financial subsidies of higher education on the part of Federal, State, and local governments. Let us identify the most significant of these factors, and also attempt to determine their rough relative significance for less affluent college aspirants.

The factor of overriding significance would seem to be massive growth of "open-door" public commuter 2-year colleges during the 1960s. In 1961 there were just over four hundred such schools enrolling about 645,000 students. By 1968 there were over seven hundred public 2-year colleges (a seventy-five per cent increase) enrolling one and three-quarters millions of students (or an increment of one hundred and seventy-one per cent in seven years).[9]

Viewed in another way, the 1960 Census Bureau follow-up of the 1959 high school seniors found about 22% of the immediate college entrants enrolling in 2-year schools. The parallel follow-up of the 1965 study in February 1967 found thirty-four per cent entering 2-year colleges, or over half again the earlier proportion. In 1967, Office of Education data indicate that about thirty-eight per cent of all first-time collegians were enrolled in 2-year colleges.[10] This latter figure includes delayed as well as immediate college entrants. It is significant to note that Fall 1966 Census data on the college population indicate that about forty-one per cent of first-time students at 2-year colleges had delayed entrance over a year following high school graduation, whereas this was the case for only twenty-seven per cent at 4-year colleges.

The significance of the 2-year versus 4-year college entrance trends becomes clear when we compare estimated costs of a year of college in the academic year 1966-1967 for the three major types of colleges. For private 4-year colleges the cost, including direct and indirect college-related expenses, was about $2,600; for public 4-year colleges the equivalent figure was $1,600; and for public 2-year colleges the cost was only about $1,100.[11]

For generally less affluent Negro youth the 2-year college appears to be particularly important. In the Far West, where such colleges are especially prevalent, just over seventy per cent of first-time Negro collegians in 1965 were attending 2-year schools, whereas this was the case for less than fifty per cent of white entrants.[12] There are relatively few Negroes in the Far West, but parallel findings pertain in states with large numbers of Negroes, and where 2-year colleges are generally available as well, such as Florida. It is also significant that in very recent years networks of public junior colleges are being legislated into existence in many Southern states with large Negro populations--states where 2-year public colleges were previously largely non-existent, such as Virginia and Alabama.[13]

---

[9] American Education, U. S. Office of Education, Washington, D. C., December-January 1968-69, p. 30.

[10] Opening Fall Enrollment in Higher Education, 1967, U. S. Office of Education, Washington, D. C., 1967, Table 2, p. 7.

[11] Froomkin et al., Students and Buildings, Office of Education, Washington, D. C., 1968, p. 14.

[12] Coleman et al., Equality of Educational Opportunity, U. S. Office of Education, OE-38001, Washington, D. C., 1966, Table 5.5.5, p. 445.

[13] For a discussion of the significance for Negroes of 2-year colleges in the South, and

Finally, the February 1967 Census Bureau follow-up yields profiles of 2- and 4-year college entrants. Two-year entrants are about a quarter again as likely as 4-year entrants to derive from blue-collar homes, half again as likely to come from families with incomes under $6,000, and once again as likely to have fathers who failed to graduate from high school.[14]

Direct student aid, principally from Federal sources, also increased a great deal during the 1960s. Federal and Federally guaranteed student loans commenced in the late 1950s, followed by Work-Study funds a few years later. Veterans' educational benefits are once again provided. Finally, the most recent major Federal program, Educational Opportunity Grants, is specifically addressed to lower socio-economic college aspirants.

We can roughly gauge the relative importance of major sources of college funding by noting that in the academic year 1966-67 students and their families supplied about fifty-seven per cent of the total outlay required to educate undergraduates; major government programs, including loans, defrayed about ten per cent of the costs; and subsidies to the institutions (including scholarship funds), passed on directly or indirectly to the students, accounted for the remaining thirty-three per cent.[15]

The Federal programs seem particularly important to the minority of less affluent students who enroll in relatively expensive and chiefly private 4-year colleges--colleges charging $500 or more per year in tuition and fees in the mid-1960s. About two in ten entrants from families with incomes under $10,000 per year chose such colleges according to the 1967 Census Bureau follow-up, and about one in ten entered colleges charging over $1,000. Students at colleges costing $500 or more were about two and a half times as likely to take out loans as students at less expensive schools.[16]

We can summarize by saying that the recent large increases in student aid would appear, at least inferentially, to make attendance at expensive private 4-year colleges possible for less affluent students. In general these are academically selective colleges. But it is the low-cost public 2-year college per se which seems to have played the principal role in the marked liberalization of access to college for less affluent students during the 1960s.

All evidence points to further expansion of 2-year colleges in the years ahead. Further increases in student aid, though far from certain, would seem probable--and especially so for less affluent racial minority youth. In short, the higher educational establishment appears to be reshaping itself to meet the needs of the "underclasses." In turn, the focus of concern is shifting from purely financial barriers to college to ones which, though correlated with social class in considerable measure, are principally non-financial in nature.

Part V. Non-Financial Determinants of College Planning[17]

Let us establish a rough ranking of better known non-financial variables associated with college entrance. We except several variables, such as the student's plans as a senior and parental aspirations for the senior, since such variables chiefly represent the net effect of various prior variables, in combination, upon the senior. However descriptively interesting, in and of themselves they tell us little about the process of reaching college.

The February 1967 Census Bureau follow-up would seem to indicate that the 2-year college not only has substantially reduced the relevance of family income to college entrance, but also has sharply reduced the relevance of several "classic" academic variables--namely, tested academic performance according to national norms and high school grades. In the 1966-67 academic year 2-year college entrants included only ten per cent more high-performance students (high-performance, that is, according to national test-score ranking) than was the case for non-entrants to college, whereas 4-year college entrants included thirty-three per cent more high-performance students than was the case for the 2-year college--or forty-three per cent more than for the non-entrants. For high school marks, non-entrants were in fact slightly more likely than 2-year entrants to include students with B or

---

especially for the relationships between 2-year colleges and primarily Negro colleges, see: Jaffe, A. J., Adams, Walter, and Meyers, Sandra G., Negro Higher Education in the 1960's, Chapter 8, Praeger Special Studies in U. S. Economic and Social Development, Frederick A. Praeger, New York, 1968.

[14] Jaffe and Adams, op. cit., footnote 1, Section IX and Part III of Appendix C (Detailed Statistical Tables).

[15] Froomkin et al., op. cit., footnote 11, p. 16.

[16] Jaffe and Adams, op. cit., footnote 1, Section VII. See also: Adams, Walter, "The Cost of College--Who Pays the Bills?", The New York Statistician, Vol. 20, No. 1, September-October 1968, pp. 3-5.

[17] For this section of the present paper, findings and detailed tabulations may be found in: Jaffe and Adams, op. cit., footnote 1. Sections VIII and IX are especially relevant, as well as Parts III and V of Appendix C (Detailed Statistical Tables). See also: Adams, Walter, "Academic Self-Image as a Strong Determinant of College Entrance and Adult Prospects: Relative Deprivation Theory Applied to High School Curriculum Choice," forthcoming article in The American Journal of Economics and Sociology.

better averages. The explanation for this apparent anomaly (in part at least) would seem to be the greater likelihood for 2-year entrants, as compared to non-entrants, to have followed the more difficult college preparatory program in company with generally more able classmates. Nevertheless, the finding remains that high school grades differentiate only between 2- and 4-year college entrants. Four-year college entrants were nearly twice as likely as 2-year entrants to have been B or better students in high school. Low ability students and students with poor grades--students who strongly tend to be lower socio-economic and racial minority ones --appear generally able to enter college by the junior college route.

It is only the high school program a student has followed which strongly differentiates 2-year college entrants from non-entrants, as well as 2-year from 4-year college entrants. Under two in ten non-entrants had followed college preparatory programs, whereas nearly six in ten 2-year entrants had done so--and over eight in ten 4-year entrants. The important point is that the high school program is entered at the start of the high school career, and for many students one program or another undoubtedly is simply assumed from a far earlier age. In brief, the financial and academic liberalization of college entrance which we have described does not appear to have altered the critical role of the early high school program decision for post-high school eventuations. Consequently, we are principally concerned with the determinants of program choice. We turn to the Coleman study data for insights regarding such determinants, and we append several illustrative and relatively detailed supporting tables.

What we found is the following:

First, entering the college preparatory program is indeed considerably related to a student's tested academic performance. It is also related to the educational tradition in the home, as evidenced by such measures as parental educational attainment. But several additional variables appear to be fully as independently related to program choice as tested ability, and even more strongly so than parental education. The student's academic self-image relative to classmates, controlling simultaneously for ability and parental education, seems to be the principal additional determinant of program choice. Academic self-image is also highly correlated with the student's perceived social status in his high school class, and we consequently assume that the academic self-image is part of a relatively unified overall assessment of self.

Next, we note that relative academic self-image, controlling for ability, is closely and positively associated with the marks a student receives in high school. Inferentially, the strength of the classroom competition determines academic self-image as much as does tested ability. Low ability students who receive good marks have generally favorable academic self-images, and vice versa.

In turn, we found that high school guidance advice, strongly related to a student's post-high school plans and his high school program, was also closely associated with the student's academic self-image and his grades. Guidance counselors appear to offer advice favoring more or less ambitious educational aspirations not alone in terms of the student's objectively tested ability, but also in terms of the student's performance relative to classmates, quite apart from ability level. For students and counselors alike the local scene often obscures clear perception of objectively tested college potential.

The relationships we have just described apply to racial majority and minority students alike, and to roughly the same extent. The fact that predominantly low ability lower socio-economic ethnic minority students plan on college about as frequently as do majority ones would appear to be attributable to the fact that minority students generally do not attend academically selective high schools. The majority and minority students base post-high school plans on identical criteria, including the relative self-image variables, but they appear to do so within rather than across ethnic lines. We may only speculate upon the effects of further and extensive desegregation of schools for the aspirations and plans of minority students.

Let us summarize the relationships we have outlined in the last few paragraphs. For racial minority and majority students alike, relative assessment of self appears to be as strong a determinant of high school curriculum, in turn the principal objective determinant of post-high school plans and behavior, as is tested ability. Both the student and the student's counselor appear to base self-image and advice as much upon the strength of the classroom competition as upon national test ranking. The net effect is that many able students fail to plan on college and that many students with marginal or even submarginal qualifications plan to enter.

Part VI. Conclusions and Implications

What we have said so far would seem to indicate considerable relevance of relative deprivation theory to high school curricular choice, and ultimately to post-high school plans and behavior. It would also seem that the high school guidance counselor is a strong mediating force in the set of complex interrelationships we have summarized. The important point is that relative self-image is a psychological variable unlikely to be significantly affected by further economic liberalization of access to college. It is a variable rooted in a student's early experience in the particular primary and secondary schools he attends. If increased proportions of able students, of whatever race and family background, are to plan on and attend the increasing roster of liberal-access colleges, it would seem imperative that students should have accurate perceptions of their own potential at least as early as the time of the pivotal high school program decision. For racial majority and minority students alike, roughly one in four who ranked above average on national ability and performance test distributions nevertheless rated themselves academically average or below average relative to classmates. Over half such students had failed to enter college preparatory curricula. Roughly two and a half times the proportion of able majority and minority seniors with unfavorable academic self-images failed to plan on college as was the case when the self-image was favorable.[18]

It would seem that extended investigation of relationships between relative academic self-image and educational plans and aspirations are in order. It would seem that the role of the guidance counselor should be determined more explicitly than is possible from available data.

Several generations ago, when high school graduation represented the passport to better jobs, high school curricular choice and its determinants were not too relevant for most students. But today, when roughly six in ten high school graduates seem destined to enter college eventually, the curricular decision is crucial. Our findings suggest that it is precisely at the time of this decision that the characterization of American education suggested by the sociologist Ralph Turner seems open to serious qualification. Comparing the British and American educational establishments, Turner spoke of the "sponsored mobility" of the former and the "contest mobility" of the latter. It would seem that the accident of more or less able classmates often limits the "openness" of the contest for American high school students. Guidance personnel, in most instances unwittingly I am sure, often reinforce erroneous student appraisals of self, and consequently limit academically reasonable post-high school aspirations.[19]

We suggest that further erosion of economic barriers to college (and many academic ones as well) will take place in the years ahead, thereby increasing the relative significance of the remaining obstacles--and principally, in our opinion, the self-image variables. National sample studies, such as the Coleman one, permit identification of such residual obstacles, and also permit some insights into their determinants. But the sheer complexity of the relationships between the relevant residual variables dictate that future liberalizing programs and policies must be based on research in depth for specific high school environments. We believe that our findings, however limited, clearly call for mounting such an effort.

---

[18] Op. cit., footnote 17.

[19] For a discussion of Turner's characerization of the American system, relative to American guidance counselling, see: Cicourel, A. V. and Kitsuse, J. I., The Educational Decision Makers, Advanced Studies in Sociology Series, The Bobbs-Merrill Company, Indianapolis, Indiana, 1963.

APPENDIX TABLES

(For additional detailed tabulations, the reader is referred to the published materials cited in the footnotes to the various sections of this paper.)

Table I. Race, Modal Level of Educational Attainment, and Per Cent Above and Below Modal Level for White Youth

Approximate year of high school graduation	White			Non-White			Ratio per cent white to per cent non-white below white modal level	Ratio per cent non-white to per cent white above white modal level
	Modal level	Per cent above white mode	Per cent below white mode	Modal level	Per cent above white mode	Per cent below white mode		
	Years	Per cent	Per cent	Years	Per cent	Per cent	Ratio	Ratio
1896 and earlier	7-8	21	37	0-6	5	86	.43	.24
1904-05	7-8	34	27	0-6	15	64	.42	.44
1914-15	7-8	44	18	0-6	18	61	.30	.41
1924-25	7-8	59	10	0-6	27	46	.22	.46
1933	12	15	51	0-6	7	80	.64	.47
1944-45	12	30	37	9-11	12	64	.58	.40
1952	12	30	30	12	17	47	.64	.57
1957	12	30	25	12	17	44	.57	.57
1962	12	35	21	12	18	44	.48	.51

Sources: Current Population Reports, Series P-20: No. 169, 1968, Table 1; No. 15, 1948, Table 1.

Table II. Proportion of Whites and Non-Whites in Recent Years Who Completed at Least the 8th, 9-11th, and 12th Grades, and the Proportion Who Completed at Least a Year of College

	Grades											
	8th			9-11th			12th			13th		
Approximate year of high school graduation	White	White	Difference White vs. Non-White	White	Non-White	Difference White vs. Non-White	White	Non-White	Difference White vs. Non-White	White	Non-White	Difference White vs. Non-White
	%	%	%	%	%	%	%	%	%	%	%	%
1935	89	61	28	75	46	29	55	26	29	19	9	10
1945	92	76	16	83	63	20	64	37	27	24	12	12
1952	94	86	8	88	79	9	70	53	17	27	17	10
1957	96	91	5	91	83	8	75	56	19	30	17	13
1962	97	93	4	93	86	7	79	56	23	35	18	17

Source: As for Table I.

Table III. Various Measures of School Retention in the United States for Cohorts of High School Graduation Age in the Late 1930s Through the Late 1960s

Year of high school graduation	High school freshmen as a proportion of fifth grade pupils four years previous	College freshmen as a proportion of fifth grade pupils eight years previous	High school graduates as proportion of high school freshmen three years previous	College freshmen as a proportion of high school graduates one year previous
	Per cent	Per cent	Per cent	Per cent
1938	77	15	54	35
1944	84	12*	48*	31*
1950	81	21	63	41
1956	86	30	67	52
1962	92	34	70	53
1967	97	40	75	55

*Drops from the 1938 proportions presumably attributable to World War II.

Source: Proportions calculated from Office of Education data in Table 181, page 126, of the Statistical Abstract of the United States, 1968.

Table IVa. Attitudes and Plans of High School Students Toward Attending College by Major Occupation Group of Head of Household, 1939, 1955, 1959

1939[a]

Occupation of head	Total	Attend college: Total	Attend college: Plan to go	Attend college: Hope to go	No college or Undecided
	%	%	%	%	%
Professional and managerial	100	75	62	13	25
Other white collar	100	70	56	14	31
Manual workers	100	47	32	15	53
Farmers and farm laborers	100	46	32	14	54
Unemployed and not in labor force	100	51	35	16	49
Total	100	54	40	14	46

1955[b]

Occupation of head	Total	Attend college: Total	Attend college: Plan to go	Interested only	No college or Undecided
	%	%	%	%	%
Professional and managerial	100	72	68	4	28
Other white collar	100	68	63	5	32
Manual workers	100	48	40	8	52
Farmers and farm laborers	100	45	38	7	55
Unemployed and not in labor force	100	50	43	7	50
Total	100	56	49	7	44

1959[c]

Professional and managerial	100	--	68	--	32
Other white collar	100	--	61	--	39
Manual workers	100	--	37	--	63
Farmers and farm laborers	100	--	34	--	66
Unemployed and not in labor force	100	--	43	--	57
Total	100	--	47	--	53

Table IVb. High School Seniors' Post-High School Plans, Fall 1959 and 1965, by Major Occupation Group of Head of Household

Major occupation group of household head	All high school seniors 1959[c]	1965[d]	Per cent change	Seniors' post-high school plans Total	College	No college	Undecided
	%	%	%	%	%	%	%
All white collar	34	37	+3				
1959				100	66	19	15
1965				100	74	18	8
Per cent change					+8	-1	-7
Manual and service	48	48	--				
1959				100	37	41	22
1965				100	52	36	12
Per cent change					+15	-5	-10
Farm	9	6	-3				
1959				100	34	39	27
1965				100	44	33	23
Per cent change					+10	-6	-4
Unemployed and not in labor force	9	9	--				
1959				100	43	37	20
1965				100	54	28	18
Per cent change					+11	-9	-2
Total	100	100					
1959				100	47	33	20
1965				100	60	29	11
Per cent change					+13	-4	-9

Table 4c. Seniors' Post-High School Plans, Fall 1959 and 1965, by Family Income

Family income	All high school seniors 1959[c]	1965[d]	Per cent change	Seniors' post-high school plans Total	College	No college	Undecided
	%	%	%	%	%	%	%
Under $3,000	19	13	-6				
1959				100	23	52	25
1965				100	46	39	15
Per cent change					+23	-13	-10
$3,000-4,999	24	17	-7				
1959				100	40	40	20
1965				100	47	38	15
Per cent change					+7	-2	-5
$5,000-7,499	28	26	-2				
1959				100	52	29	19
1965				100	58	31	11
Per cent change					+6	+2	-8
$7,500 and over	29	44	+15				
1959				100	68	17	15
1965				100	71	22	7
Per cent change					+3	+5	-8
Total	100	100					
1959				100	49	32	19
1965				100	60	29	11
Per cent change					+11	-3	-8

Table 4d. Comparison of High School Seniors' Post-High School Plans, Fall 1959 and 1965, by Family Income Roughly Adjusted for Changes in Income Distribution

Adjusted family income 1959[c]	1965[d]	Year	Per cent planning on college
Under $3,000	Under $4,000		
		1959	23
		1965	46
Per cent change			+23
$3,000-4,999	$4,000-5,999		
		1959	40
		1965	52
Per cent change			+12
$5,000-7,499	$6,000-8,499		
		1959	52
		1965	65
Per cent change			+13
$7,500 and over	$8,500 and over		
		1959	68
		1965	74
Per cent change			+6

Sources for Table IV.

[a]Source: Unpublished data from a 1939 survey, conducted by Elmo Roper and Associates, Roper Commercial number 15. The 1,148 respondents were a national sample of persons under 20 years of age. Those already in college are excluded from this tabulation. The specific question was: "Do you plan on going to college?" The answers indicating attitudes favorable to attending college were: "Plan on going" and "Hope to go." The totals for all occupations were obtained by weighting the replies for each occupation by the distribution for total U.S., 1940, from 1940 Census of Population, Families: Employment Status, Table 19, Distribution of males aged 35-44 having children under 18 years of age.

[b]Source: Unpublished data derived from a study conducted by the Educational Testing Service in Spring 1955, Background Factors Relating to College Plans and College Enrollment Among Public High School Students. The 35,400 respondents were a national sample. The specific questions were: "What is your father's occupation? What does he do . . . ?" and, in relation to college plans, "Think of what you would really like to do when you finish high school . . ." and "What do you really think you will do when you finish high school?" To the latter two questions, answers favorable to going to college, without expectation of so doing, were one answer category, expectations of attending formed another, and negative answers to both queries formed a third category. The respondents also specified, if they intended to enter college, whether they planned to do this immediately or later, after a period of work. For comparability with the 1939 study above, those foreseeing a delay in entrance were not included in the "plan on going to college" category.

[c]Source: Data derived from a national survey, Educational Status, College Plans, and Occupational Status of Farm and Nonfarm Youths: October 1959, by James C. Cowhig and Charles B. Nam, U.S. Bureau of the Census, Series ERS (P-27), No. 30, August 1961. The study sample was composed of 1,279 high school seniors dwelling in the approximately 35,000 households interviewed in connection with the monthly population sample survey of the Bureau of the Census. The specific question was: "Does . . . plan to attend college next fall?"

[d]Source: Data derived from a Census Bureau study, paralleling the one described in c above. The appreciably larger number of households sampled in 1965 yielded 1,464 high school seniors for tabulation.

Table V. 1966 High School Graduates Entering and Not Entering College the Following Fall or Early Winter, and For Those Who Entered, the Type of College Entered

Student characteristics	Did not enter college	Entered a 2-year college	Entered a 4-year college	All college entrants	All high school graduates
	%	%	%	%	%
Sex:					
Male	46	58	54	55	50
Female	54	42	46	45	50
Both sexes	100	100	100	100	100
Age, October, 1966					
18 years or less	75	86	95	92	83
19 years or more	25	14	5	8	17
All ages	100	100	100	100	100
Family income:					
Under $3,000	16	5	5	5	11
$3,000-3,999	9	6	4	5	7
$4,000-5,999	24	22	13	16	20
$6,000-7,499	17	11	15	14	15
$7,500-9,999	18	25	21	22	20
$10,000-14,999	14	23	28	26	20
$15,000 and over	2	8	14	12	7
All incomes	100	100	100	100	100
Under $7,500	66	44	37	39	53
$7,500 and over	34	56	63	61	47
All incomes	100	100	100	100	100
Under $6,000	49	33	22	26	38
$6,000 and over	51	67	78	74	62
All incomes	100	100	100	100	100
Occupation, head of household:					
Blue collar	75	56	44	48	62
White collar	25	44	56	52	38
All occupations	100	100	100	100	100
Father's education:					
11 grades or less	60	44	23	30	45
12 grades	27	27	37	34	30
13-15 grades	9	21	14	16	13
16 grades or more	4	8	26	20	12
All levels	100	100	100	100	100
11 grades or less	60	44	23	30	45
12 grades or more	40	56	77	70	55
All levels	100	100	100	100	100
11 grades or less	--	50	50	100	--
12 grades	--	27	73	100	--
13-15 grades	--	43	57	100	--
16 grades or more	--	14	86	100	--
Ability score:					
High	19	29	62	51	35
Medium and low	81	71	38	49	65
All levels	100	100	100	100	100
Average high school mark:					
B- or better	45	39	73	61	53
C+ or poorer	55	61	27	39	47
All marks	100	100	100	100	100

Student characteristics	Did not enter college	Entered a 2-year college	Entered a 4-year college	All college entrants	All high school graduates
	%	%	%	%	%
High school curriculum:					
College preparatory	19	56	84	74	45
All other	81	44	16	26	55
All curricula	100	100	100	100	100
College plans as high school senior:					
No college	66	17	8	11	40
2-year college only	15	25	2	10	12
4-year college only	11	20	81	60	34
2 + 4 year college**	8	38	9	19	14
All plans	100	100	100	100	100
No college	87	7	6	13	100
2-year college only	63	33	4	37	100
4-year college only	17	9	74	83	100
2 + 4 year college**	34	45	21	66	100
All plans	53	16	31	47	100
No college	--	52	48	100	--
2-year college only	--	90	10	100	--
4-year college only	--	11	89	100	--
2 + 4 year college**	--	68	32	100	--
All plans	--	34	66	100	--
Number of cases*	1,387,696	419,268	805,549	1,224,817	2,612,513

*National sample inflated to national totals, according to known national distributions for the major demographic variables.

**"2 + 4-year college" designates students who intended to enter a junior college initially, and subsequently transfer to a senior one.

Source: Unpublished data from February 1967 Census Bureau follow-up of 1965-66 high school seniors.

Table VI. College Financing of 1965-66 High School Seniors Entering College Immediately in Fall 1966 and Early 1967

Proportion of college expenses met by students' families by family income

Income	Over 75%	75% and under	Total
	%	%	%
Family income of student, total	54	46	100
Under $5,000	41	59	100
$5,000 to $9,999	47	53	100
$10,000 to $14,999	59	41	100
$15,000 and over	74	26	100

Proportion of college expenses met by students' families by supplementary sources of financing

Non-family sources of financing	Over 75%	75% and under	Total
Loan	25%	29%	27%
Summer earnings	64	63	63
Other savings	28	22	24
Scholarship	14	30	24
Veterans' benefits	--	2	1
Employment during school year	11	34	25
College	5	13	10
Non-college	6	21	15
All further sources	10	9	10
Total	152*	189*	174*
Per cent	39	61	100

*Percentages add to more than 100 because of multiple mentions.

College tuition and fees

Family income	Under $250	$250-499	$500-999	$1,000 and over	Total
	%	%	%	%	%
Under $10,000	49	31	11	9	100
$10,000 and over	30	36	13	21	100
Total	41	33	12	14	100

Proportion of expenses met by families of students in private and public colleges

Auspice of college attended and student's residence while attending

	Public college			Private college		
Proportion of college expenses met by family*	Family or relatives*	Dormitory, frat house, rooming house, etc.	All living arrangements	Family or relatives*	Dormitory, frat house, rooming house, etc.	All living arrangements
	%	%	%	%	%	%
More than 75%	57	62	59	62	64	64
50 to 75%	16	14	15	19	14	15
Some, but less than 50%	27	12	20	19	17	18
None	--	12	6	--	5	3
Total	100	100	100	100	100	100

*Our data indicate that nearly all students who live at home do so at family expense, and do not report this item as part of college costs. The table includes this imputed expenditure.

[Table VI. continued]

Source of financing by college tuition level

Non-family sources of financing	College tuition and fees: Under $500	$500 and over	Total*
	%	%	%
Loan	16	39	23
Summer earnings	65	71	67
Other savings	24	27	25
Scholarship	22	35	25
Veterans' benefits	2	--	1
Employment during school year:	26	22	25
College	8	13	9
Non-college	18	9	16
All further sources	10	8	9
Total	165**	203**	175**

*Proportions citing various sources in this table differ slightly from proportions in the earlier table (where family financing is the independent variable) because of differences in "non-response" for the independent variables. The differences are too slight to affect any of the findings which we report.

**Sources add to more than 100% because of multiple mentions.

Data source: Unpublished data from 1967 Census Bureau follow-up of 1965-66 high school seniors.

Table VII.

Proportion of 1965-1966 high school seniors planning on college by high school curriculum: Coleman data – College preparatory	Coleman data – Other programs	Census data – College preparatory	Census data – Other programs	Proportion of 1965-1966 high school graduates entering college, as of February 1967, by high school curriculum: Census data – College preparatory	Census data – Other programs
90	46	90	47	78	22

Ratio college entrants to college planners* by high school curriculum

Census data

College preparatory	Other programs
.87	.47

*This is not precisely a measure of proportions of college planners who in fact entered college, since a small proportion of the entrants (about one in nine) had been non-planners.

Table VIII. Senior's Estimate of Own Brightness Relative to Classmates and Senior's National Verbal Ability Ranking by Senior's High School Curriculum

Majority Seniors

Senior's verbal ability and high school curriculum	Total	Senior's estimate of relative brightness: Very low	Low to average	Above average	Among brightest	Don't know
Low ability						
Number reporting	5,716	655	3,818	775	210	258
Total	100%	100%	100%	100%	100%	100%
College prep	9	4	8	16	21	10
General	32	38	33	26	28	29
Other courses	59	58	59	58	51	61
Medium ability						
Number reporting	28,456	1,391	17,442	7,842	1,230	551
Total	100%	100%	100%	100%	100%	100%
College prep	30	14	23	46	44	23
General	26	42	29	19	20	30
Other courses	44	44	48	35	36	47
High ability						
Number reporting	28,587	299	6,991	14,910	5,968	419
Total	100%	100%	100%	100%	100%	100%
College prep	71	44	50	75	87	68
General	14	31	23	11	7	16
Other courses	15	25	27	14	6	16
All ability levels						
Number reporting	62,759	2,345	28,251	23,527	7,408	1,228
Total	100%	100%	100%	100%	100%	100%
College prep	47	15	27	64	78	36
General	21	39	28	14	10	25
Other courses	32	46	45	22	12	39

Source: Our own tabulations of the Coleman Study 12th grade data.

[Table VIII. continued]

Minority Seniors

Senior's verbal ability and high school curriculum	Total	Senior's estimate of relative brightness: Very low	Low to average	Above average	Among brightest	Don't know
Low ability						
Number reporting	14,631	815	7,776	3,096	1,418	1,526
Total	100%	100%	100%	100%	100%	100%
College prep	15	8	12	20	22	13
General	25	32	27	23	23	24
Other courses	60	60	61	57	55	63
Medium ability						
Number reporting	12,386	367	6,095	3,663	1,605	656
Total	100%	100%	100%	100%	100%	100%
College prep	35	17	29	44	44	29
General	26	35	27	21	27	30
Other courses	39	48	44	35	29	41
High ability						
Number reporting	3,091	78	756	1,369	792	96
Total	100%	100%	100%	100%	100%	100%
College prep	67	29	55	73	73	61
General	16	27	21	14	12	17
Other courses	17	44	24	13	15	22
All ability levels						
Number reporting	30,108	1,260	14,627	8,128	3,815	2,278
Total	100%	100%	100%	100%	100%	100%
College prep	28	12	21	40	42	19
General	25	33	27	20	22	26
Other courses	47	55	52	40	36	55

Source: Our own tabulations of the Coleman Study 12th grade data.

Table IX. Senior's Estimate of Own Brightness Relative to Classmates, Verbal Ability of Senior, and Highest Grade Completed by Senior's Mother by Post-High School Plans of Senior

Majority seniors

Post-high school plans of senior	Above average ability					Low to average ability					Very low ability					
	13 or more	12	9-11	8 or less & don't know	All grades	13 or more	12	9-11	8 or less & don't know	All grades	13 or more	12	9-11	8 or less & don't know	All grades	All Abilities
	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%
Estimated brightness above average																
No college	4	12	24	24	13	12	24	39	39	28	25	40	50	50	44	18
College probably	11	18	25	27	18	19	26	30	32	27	29	32	29	31	30	21
College definitely	85	70	51	49	69	69	50	31	29	45	56	28	21	19	26	61
All plans	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Per cent of cases	9.0	17.0	4.9	3.0	33.9	2.2	6.7	3.5	2.4	14.8	0.2	0.5	0.4	0.4	1.5	50.2
Estimated brightness average or below																
No college	16	29	45	48	34	24	44	58	63	50	34	60	68	71	65	47
College probably	29	31	28	27	29	32	31	27	25	29	33	27	24	23	25	28
College definitely	55	40	27	25	37	44	25	15	12	21	33	13	8	6	10	25
All plans	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Per cent of cases	1.7	5.8	2.6	1.7	11.8	2.6	13.0	8.5	6.6	30.7	0.3	2.3	2.3	2.4	7.3	49.8
Per cent whose estimated brightness is average or below	16	25	35	36	26	54	66	71	69	68	67	81	84	85	82	50

Source: Our own tabulations of the Coleman Study 12th grade data.

[continued]

[Table IX. continued]

Minority seniors

Post-high school plans of senior	Above average ability: 13 or more	12	9-11	8 or less & don't know	All grades	Low to average ability: 13 or more	12	9-11	8 or less & don't know	All grades	Very low ability: 13 or more	12	9-11	8 or less & don't know	All grades	All Abilities
	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%
Estimated brightness above average																
No college	6	10	14	22	12	10	15	18	24	17	13	21	25	34	27	20
College probably	14	24	33	25	23	21	31	35	38	33	24	39	42	41	39	34
College definitely	80	66	53	53	65	69	54	47	38	50	63	40	33	25	34	46
All plans	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Per cent of cases	2.0	2.7	1.5	1.4	7.6	2.6	5.4	5.6	4.9	18.5	1.3	3.4	5.6	6.3	16.6	42.7
Estimated brightness average or below																
No college	12	28	30	44	29	16	29	36	45	35	22	33	42	49	42	39
College probably	22	28	36	30	29	32	38	38	38	38	38	41	40	38	39	37
College definitely	66	44	34	26	42	52	33	26	17	27	40	26	18	13	19	24
All plans	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Per cent of cases	0.6	1.0	0.7	0.8	3.1	2.1	6.5	7.0	7.3	22.9	1.4	5.8	10.6	13.5	31.3	57.3
Per cent whose estimated brightness is average or below	24	28	31	35	29	45	55	56	66	55	53	63	66	68	65	57

Source: Our own tabulations of the Coleman Study 12th grade data.

Table X. Post-High School Plans of High School Senior by Senior's Tested National Verbal Ability Standing, by Senior's Own Estimate of His Brightness Relative to Classmates, and by Senior's High School Grades

Racial majority males

Senior's verbal ability, estimated brightness relative to classmates, and high school grades	Post-high school plans of high school senior: College definitely %	College probably %	No college %	All plans %
Above average verbal ability				
Above average brightness				
Grades: A or B	79	16	5	100
C	52	30	18	100
D	27	32	41	100
All grades	71	20	9	100
Average or below brightness				
Grades: A or B	52	30	18	100
C	35	35	30	100
D	15	30	55	100
All grades	38	34	28	100
All levels of brightness				
Grades: A or B	76	17	7	100
C	44	32	24	100
D	20	30	50	100
All grades	63	23	14	100
Low to average verbal ability				
Above average brightness				
Grades: A or B	57	27	16	100
C	39	33	28	100
D	15	30	55	100
All grades	48	30	22	100
Average or below brightness				
Grades: A or B	34	34	32	100
C	22	34	44	100
D	12	25	63	100
All grades	24	33	43	100
All levels of brightness				
Grades: A or B	49	29	22	100
C	27	34	39	100
D	13	25	62	100
All grades	33	32	35	100

Source: Our own tabulations of the Coleman Study 12th grade data.

[Table X. continued]

Racial majority males

Senior's verbal ability, estimated brightness relative to classmates, and high school grades		Post-high school plans of high school senior: College definitely %	College probably %	No college %	All plans %
Very low verbal ability					
Above average brightness					
Grades:	A or B	39	35	26	100
	C	24	31	45	100
	D	17	20	63	100
	All grades	29	31	40	100
Average or below brightness					
Grades:	A or B	23	32	45	100
	C	12	31	57	100
	D	5	20	75	100
	All grades	13	30	57	100
All levels of brightness					
Grades:	A or B	30	33	37	100
	C	14	31	55	100
	D	7	20	73	100
	All grades	17	30	53	100
All levels of verbal ability					
Above average brightness					
Grades:	A or B	73	19	8	100
	C	44	32	24	100
	D	20	29	51	100
	All grades	63	23	14	100
Average or below brightness					
Grades:	A or B	39	32	29	100
	C	24	34	42	100
	D	11	24	65	100
	All grades	26	33	41	100
All levels of brightness					
Grades:	A or B	66	21	13	100
	C	31	33	36	100
	D	13	25	62	100
	All grades	46	27	27	100

Source: Our own tabulations of the Coleman Study 12th grade data.

[Table X. continued]

Racial minority males

Senior's verbal ability, estimated brightness relative to classmates, and high school grades	Post-high school plans of high school senior			
	College definitely	College probably	No college	All plans
	%	%	%	%
Above average verbal ability				
Above average brightness				
Grades: A or B	71	21	8	100
C	48	33	19	100
D	32	32	36	100
All grades	63	25	12	100
Average or below brightness				
Grades: A or B	42	40	18	100
C	36	35	29	100
D	16	19	65	100
All grades	36	35	29	100
All levels of brightness				
Grades: A or B	67	24	9	100
C	43	34	23	100
D	23	24	53	100
All grades	55	28	17	100
Low to average verbal ability				
Above average brightness				
Grades: A or B	54	33	13	100
C	38	43	19	100
D	25	25	50	100
All grades	47	37	16	100
Average or below brightness				
Grades: A or B	33	42	25	100
C	23	42	35	100
D	12	31	57	100
All grades	25	41	34	100
All levels of brightness				
Grades: A or B	47	36	17	100
C	28	43	29	100
D	15	29	56	100
All grades	35	39	26	100

Source: Our own tabulations of the Coleman Study 12th grade data.

[Table X. continued]

Racial minority males

Senior's verbal ability, estimated brightness relative to classmates, and high school grades	College definitely	College probably	No college	All plans
	%	%	%	%
Very low verbal ability				
Above average brightness				
Grades: A or B	38	39	23	100
C	23	45	32	100
D	21	36	43	100
All grades	30	42	28	100
Average or below brightness				
Grades: A or B	26	43	31	100
C	16	42	42	100
D	12	32	56	100
All grades	19	41	40	100
All levels of brightness				
Grades: A or B	33	41	26	100
C	18	43	39	100
D	14	33	53	100
All grades	23	42	35	100
All levels of verbal ability				
Above average brightness				
Grades: A or B	52	33	15	100
C	33	42	25	100
D	24	32	44	100
All grades	44	36	20	100
Average or below brightness				
Grades: A or B	31	42	27	100
C	20	42	38	100
D	12	30	58	100
All grades	22	41	37	100
All levels of brightness				
Grades: A or B	46	36	19	100
C	24	42	34	100
D	15	31	54	100
All grades	32	39	29	100

Source: Our own tabulations of the Coleman Study 12th grade data.

Table XI. Senior's Estimate of His Own Brightness Relative to Classmates and His Tested Level of Verbal Ability by the Post-High School Guidance Advice Offered the Senior

Post high school graduation guidance offered to senior	Among brightest				Above average				Average			
	Very low	Low to average	Above average	All levels	Very low	Low to average	Above average	All levels	Very low	Low to average	Above average	All levels
	%	%	%	%	%	%	%	%	%	%	%	%
Majority males												
College	40	63	89	83	44	64	81	74	29	41	58	44
No college	60	37	11	17	56	36	19	26	71	59	42	56
All advice	100	100	100	100	100	100	100	100	100	100	100	100
Number of cases	131	775	3,273	4,179	438	3,933	7,119	11,490	1,666	7,368	3,190	12,224
Minority males												
College	56	72	80	68	54	66	78	64	44	49	62	47
No college	44	28	20	32	46	34	22	36	56	51	38	53
All advice	100	100	100	100	100	100	100	100	100	100	100	100
Number of cases	619	695	362	1,676	1,297	1,567	653	3,517	2,869	2,251	348	5,468
Majority females												
College	40	59	79	76	29	54	73	66	18	29	42	31
No college	60	41	21	24	71	46	27	34	82	71	58	69
All advice	100	100	100	100	100	100	100	100	100	100	100	100
Number of cases	60	358	2,755	3,173	273	3,413	7,470	11,156	1,694	8,525	4,067	14,286
Minority females												
College	61	72	84	71	54	64	78	63	38	46	54	42
No college	39	28	16	29	46	36	22	37	62	54	46	58
All advice	100	100	100	100	100	100	100	100	100	100	100	100
Number of cases	631	784	384	1,799	1,426	1,809	664	3,899	3,869	3,279	340	7,488

Source: Our own tabulations of the Coleman Study 12th grade data.

[Table XI. continued]

Post high school graduation guidance offered to senior	Below average				All estimates			
	Very low	Low to average	Above average	All levels	Very low	Low to average	Above average	All levels
	%	%	%	%	%	%	%	%
Majority males								
College	18	24	40	25	31	48	77	61
No college	82	76	60	75	69	52	23	39
All advice	100	100	100	100	100	100	100	100
Number of cases	314	715	192	1,221	2,549	12,791	13,774	29,114
Minority males								
College	27	31	32	29	47	57	73	55
No college	73	69	68	71	53	43	27	45
All advice	100	100	100	100	100	100	100	100
Number of cases	345	170	44	559	5,130	4,683	1,407	11,220
Majority females								
College	8	19	33	17	19	36	65	49
No college	92	81	67	83	81	64	35	51
All advice	100	100	100	100	100	100	100	100
Number of cases	256	553	86	895	2,283	12,849	14,378	29,510
Minority females								
College	18	29	25	22	43	54	73	51
No college	82	71	75	78	57	46	27	49
All advice	100	100	100	100	100	100	100	100
Number of cases	349	143	20	512	6,275	6,015	1,408	13,698

Source: Our own tabulations of the Coleman Study 12th grade data.

A COMPUTER MODEL TO MEASURE THE REQUIREMENTS FOR STUDENT AID IN HIGHER EDUCATION

Joseph Froomkin

Murray Pfeferman

Office of Program Planning and Evaluation
Office of Education
Department of Health Education and Welfare

In the Spring of 1969 the Office of Program Planning and Evaluation of the Office of Education undertook the task to estimate the Federal resources which will be required to fulfill the increasing aspiration of young Americans for Post-Secondary Education.

Changes in Aspirations in Post-Secondary Education

While college attendance increased proportionately for all income groups between 1940 and 1960, patterns of desire to attend college have changed dramatically between '60 and '69.

> Between 1939 and 1959, young people from all income groups increased their aspirations to attend college at a uniform rate.
>
> Between 1960 and 1966, a new trend started manifesting itself. The aspirations of the poor for a college degree began to catch up with those of the rich. Twice as high a proportion of high school seniors from the lowest income quartile hoped to attend college in 1966 as did in 1959. The increase was from 23 percent to 46 percent. The number of children from families in the second income quartile – families whose income is still below the median – who expected to enroll in college rose from 40 percent in 1959 to 52 percent in 1966. This was an increase of 30 percent. The desire to attend college grew more modestly in the upper two income quartiles, from 52 percent to 65 percent of seniors in the third quartile, and from 68 percent to 74 percent of those in the highest quartile in 1966.

These changes in expectations are reflected in enrollments. 230,000 more freshmen enrolled in college full time in the Fall of 1968, than would have been expected if the trend of 1956-65 had been followed. About 65,000 of these freshmen came from families with incomes of less than $5,000 a year, the bottom quartile of the income distribution. Another 61,000 came from the second quartile. Altogether, 217,000 out of the quarter million total increase in enrollments came from other than the traditional sources of college students in the 1950's.

The increasing rate of post-secondary school attendance by students from poorer families became apparent soon after the enactment of the Higher Education Act of 1965. During the academic year that began in 1966, some 900,000 students received financial assistance under one or more of the Federal aid programs administered by the U.S. Office of Education. During 1968-69, the number of young people aided by Office of Education programs alone is expected to exceed 1.5 million students. Meanwhile, considerable additional aid is also available through the Veterans' Administration. The Veterans' Administration will have contributed $323 million to student finances in the academic year 1968-69, and expects to increase this aid to $425 million in the academic year starting in September 1969.

Considerations in Projecting Demand for College Attendance

The key factors which will affect college enrollments in the next few years are: (1) the propensity of high school graduates with different levels of academic achievement and financial resources to enroll in college, (2) the time schedule by which they enroll, i.e., immediately after high school or a number of years later, (3) the persistence rates of different types of students, and (4) the availability of student aid to make their desires come true.

The model takes into account the first three factors, and ignores the fourth. Its purpose is to project the number of students who may wish to attend and, to estimate the demand for student aid.

The estimates of the proportion of freshmen enrolling in colleges are computed taking into account dynamic change in enrollment rates in contrast to more conventional level projections. First-time enrollments by income class vary from year to year according to past trends in the propensities to enroll in college. Thus, the proportion of high school graduates from families in the lowest income quartile in proportion to all seniors who will enroll in college one year after high school graduation is projected to increase to 15 percent in 1976, in comparison with the rate of 10 percent observed by Project Talent for 1960, and an estimated entry rate of 12 percent in the Fall of 1968.

Estimates of attrition rates by ability and income group were also derived from information which has become available from special tabulations of the five-year follow-up interviews of Project Talent. These estimates were checked, and adjusted whenever necessary, to conform with U.S. Bureau of the Census estimates of attendance by age. Together with more refined assumptions about the timing of entry and enrollment rates by income group, the estimates of attrition rates present a much more realistic representation of the social demand for education at the post-secondary level, than past projections.

Varying proportions of children from families with different levels of affluence enroll in the year after graduation, and also delay enrollment at different rates. Judging from the five-year follow-up data of Project Talent, as well as follow-up of high school seniors conducted by the U.S. Bureau of the Census in 1959 and 1966, the pattern of delayed enrollments has remained fairly constant over the past few years. Roughly nine out of ten children whose parents are in the top quartile of the population in terms of income are likely to enroll in college in the year following high school graduation. By contrast, only five out of ten children whose families are in the lowest income quartile are likely to enroll in the same year. These tendencies are included in the model.

The basic statistics of high school graduates were taken from the National Center for Educational Statistics series. The freshman class of 1976 is already born, and barring some unforeseen changes in the trend of dropouts from high school, the estimates of high school graduates by year and by sex prepared by the National Center for Educational Statistics, U.S. Office of Education, are likely to be fairly accurate.

Test of the Model in Forecasting Past Enrollments

A test of the predictive accuracy of the estimates underlying the calculations of the model is to compare them with estimates of previous years enrollments presented by the Office of Education. For the period 1960-68, the estimates of the model when compared with those of the National Center never deviate more than 2.5 percent. (See Table 3)

Given the large number of factors taken into consideration in deriving the model, it is highly encouraging that the projections of past

TABLE 1

DEGREE CREDIT FULL-TIME EQUIVALENT ENROLLMENT – OPPE MODEL
BY INCOME QUARTILE
(thousands of students)of

	Low	2nd	3rd	High	Total
1960	335.68	513.64	842.77	1289.77	2981.89
1961	379.58	563.98	916.56	1386.801	3246.93
1962	427.24	612.63	983.77	1472.85	3496.49
1963	478.24	659.65	1047.52	1552.12	3737.53
1964	541.69	723.16	1135.36	1660.15	4060.37
1965	623.37	808.09	1252.57	1805.48	4489.52
1966	712.07	894.38	1368.17	1950.01	4924.63
1967	806.44	981.98	1484.92	2101.56	5374.90
1968	895.90	1062.05	1587.22	2235.66	5780.82
1969	974.77	1131.23	1670.02	2337.54	6113.56
1970	1050.06	1198.66	1755.15	2436.09	6493.97
1971	1128.74	1273.17	1863.63	2569.72	6835.26
1972	1209.03	1350.70	1984.29	2725.39	7269.41
1973	1278.64	1417.82	2084.70	2849.02	7630.18
1974	1339.70	1476.91	2169.55	2947.67	7933.83
1975	1395.04	1530.35	2245.95	3039.32	8210.66
1976	1446.19	1579.32	2320.38	3131.92	8477.81
1977	1493.66	1624.88	2398.87	3225.50	8737.92

Source: Office of Program Planning and Evaluation, U.S. Office of Education Model (OPPE Model).

TABLE 2

DEGREE CREDIT FULL-TIME ENROLLMENT – COMPLETE EGALITY MODEL
BY INCOME QUARTILE
(thousands of students)

Income quartiles	Low	2nd	3rd	High	Total
1970	1129.0	1582.5	2009.5	2436.1	7157.1
1971	1188.8	1667.7	2118.9	2569.7	7545.1
1972	1254.6	1764.3	2244.9	2725.4	7989.2
1973	1311.7	1844.4	2346.8	2849.0	8352.0
1974	1362.4	1912.1	2430.1	2947.7	8652.2
1975	1410.3	1975.6	2507.7	3039.3	3932.9
1976	1456.0	2037.8	2585.1	3131.9	9210.9
1977	1500.8	2099.6	2662.9	3225.5	9288.8

Source: OPPE Model.

TABLE 3

TOTAL ENROLLMENT – OPPE MODEL, BY INCOME QUARTILE AND TOTAL

(thousands of students)

	Low	2nd	3rd	High	Total	OE[1]	Difference	Percent
1960	419.0	624.9	1011.5	1529.7	3585.0	3583.0	2.0	.06
1961	474.3	686.7	1100.4	1645.2	3906.6	3861.0	45.6	1.17
1962	534.4	746.4	1181.2	1747.4	4209.3	4175.0	34.3	.81
1963	599.1	804.2	1257.9	1841.3	4502.4	4495.0	7.4	.16
1964	680.3	882.8	1364.2	1970.2	4897.6	4950.0	-52.4	-1.07
1965	785.6	988.5	1506.3	2143.7	5424.1	5526.0	-101.9	-1.88
1966	900.0	1095.8	1646.3	2315.7	5957.8	5885.0	72.8	1.22
1967	1021.8	1204.8	1787.5	2495.2	6509.4	6348.0	161.4	2.48
1968	1137.7	1204.6	1911.6	2654.3	7008.4	6983.0 [2]	25.4	0.04
1969	1240.7	1391.4	2012.8	2775.9	7420.7			
1970	1339.4	1476.0	2116.8	2893.7	7825.9			
1971	1442.5	1569.4	2248.7	3052.2	8312.7			
1972	1547.4	1666.2	2395.2	3236.4	8845.3			
1973	1638.6	1750.3	2517.6	3383.3	9289.9			
1974	1718.8	1824.4	2621.2	3501.4	9665.7			
1975	1791.5	1891.3	2714.5	3611.1	10008.6			
1976	1858.5	1952.6	2805.4	3721.6	10338.0			

1 National Center for Educational Statistics, *Projections of Educational Statistics to 1977-78,* (Washington, D.C.: Government Printing Office), p. 16 (all except 1968).

2 National Center for Educational Statistics, *Opening Fall Enrollment in Higher Education, 1968,* Washington, D.C.: Government Printing Office), p. 6.

Source: All data except as noted in (1) OPPE model.

experience and estimates made on an independent basis are so close together.

Total Enrollment Projections

The projection of total enrollments derived by the model appear in Table 3. It is estimated that total enrollment may exceed ten million students by 1976, nearly a million students above the NCES estimate of total enrollments. The difference between the two estimates is due to the difference in the methods used to derive them and, more importantly, the purpose for which the estimates are constructed. The NCES projects past trends, while the model presented here predicts future demand for higher education. A word of caution should be injected. Both the model and NCES projections may fail to forecast accurately if drastic shifts in attitudes or economic conditions occur in the next few years.

The projections of enrollment by income quartile indicate that despite the higher first time entry rates into college of lower income students, the participation of lower income students in the total population of higher education is not likely to change drastically in the next few years, with the lowest half of the distribution picking up a few percentage points in the total enrollment moving from 29 percent in 1960, and 35 percent in 1968 to 37 percent in 1976.

Full-Time Equivalent Enrollment Estimates

Total enrollment is an approximate measure of the load placed upon institutions of higher education. The more conventional measure of the burden is the full-time equivalent enrollment for 1976, reproduced in Table 1, of 8.5 million students for the OPPE model. The model projects equal growth in numbers of full-time equivalent students between 1960 and 1968, and 1968 and 1976, but a decline in the rate of growth.[1] The increases of 46 percent for the period 1968 to 1976 is below the 90 percent increase between 1960 and 1968.

Estimates of Graduate and Undergraduate Enrollments

Further estimates were made to separate total enrollment into undergraduate and graduate enrollment, and into full-time enrollment (see Table 4). Different rates of full-time attendance were imputed to each income/achievement quartile. These rates, derived from Project Talent one-year follow-up data were adjusted to conform with observations for the total population collected by NCES in 1964.

There is substantial difference in the proportion of full-time students by income quartile. In the lowest income quartile, it is estimated that only 64 percent of the students attend full time; 90 percent of students in the higher income quartile attend full time.

The number of undergraduate and graduate students by income quartile was estimated on the basis of the trends (1) derived from NCES of the proportion of graduate to undergraduate students, and (2) the estimated number of graduate students. The estimate of expected graduate students by income quartile is based on aspirations for graduate degrees of the Project Talent population. These were compared with data on social origins available in a study of graduate students conducted by NCES in 1965.[2]

The resulting estimates of total graduate and undergraduate students were adjusted for differences in full-time and part-time attendance by income quartiles. It was assumed that income differentials which affected full-time attendance of undergraduates would also apply to graduate students. These factors were applied to the much lower full-time attendance patterns of graduate students.

The resulting estimates of total undergraduate and graduate enrollments, and the number of full-time undergraduate and graduate students appears in Table 4.

The "Complete Equality" Projection

An alternative projection which, from 1970 on, ascribed propensities to enroll high school seniors and retention rates of college students from families in the highest income quartile is reproduced in Table 5. It estimates total enrollments of 11.2 million in 1976; 870,000 more than the enrollments produced by the OPPE model. If these patterns are followed, enrollment of children from the lowest half of the income distributions could constitute some 40 percent of the total enrollment.

TABLE 5

DEGREE CREDIT TOTAL ENROLLMENT – COMPLETE EGALITY MODEL BY INCOME QUARTILE
(thousands of students)

	Low	2nd	3rd	High	Total
1970	1129.04	1582.48	2009.46	2436.09	7157.06
1971	1188.77	1667.70	2118.87	2569.72	7545.05
1972	1254.61	1764.28	2244.93	2725.39	7989.21
1973	1311.70	1844.44	2346.83	2849.02	8351.99
1974	1362.35	1912.09	2430.05	2947.67	8652.17
1975	1410.04	1975.57	2507.67	3039.32	8932.84
1976	1456.04	2037.78	2585.12	3131.92	9210.86
1977	1500.84	2099.62	2662.85	3225.50	9488.82

Source: OPPE Model.

The assumption underlying this projection assumes that in 1976 52.5 percent of the graduating class will enroll in college the year following graduation as compared to a little over 40 percent in 1968. The dropout rate is also reduced drastically, with the assumption made that 66 percent of the entering class will receive a B.A. within five years of high school graduation, as compared to about 50 percent, the rate observed in the late 1960's. The figures presented in Table 5 can thus be taken as an upper limit of possible enrollments under conditions of availability of adequate student aid and drastic shifts in attitudes towards college attendance.

The reason why children from families in the lower income quartiles are projected to attend only at .8 the rate of children of those in the upper quartile is their poorer high school records. Since propensity to persist in one's post-secondary education is directly related to high school performance, even with the removal of financial constraints, some inequalities still remain.

The Effect of Financial Limitations on Attendance Patterns

A comparison of projected total enrollment, full-time equivalent enrollment, and full-time enrollment highlights the impact of the difference in financial circumstances on attendance patterns. Table 6 shows that the relationship of total enrollment between the demand projection and the complete equality projection is slightly over 8 percent for 1976. Full-time equivalent enrollment increases also by 8 percent, and full-time enrollment by 9 percent.

TABLE 6

RELATIONSHIP OF COMPLETE EGALITY TO OPPE MODEL
(Complete Egality as a Percent of OPPE Model)

	Total enrollment Percent	Full-time enrollment Percent	Full-time equivalent Percent
1970	111.15	111.11	111.13
1971	111.65	110.43	110.38
1972	109.74	110.01	109.90
1973	109.24	109.61	109.46
1974	108.80	109.22	109.04
1975	108.50	108.90	108.79
1976	108.33	108.85	108.64

Source: OPPE Model.

The removal of all financial constraints has only a moderate effect on the number of students likely to attend institutions of higher education. By contrast, the effect is somewhat pronounced for students who are likely to attend full time. In other words, if the assumptions underlying the projections are correct, the removal of financial constraints is likely to have a more pronounced effect upon the intensity of studies than they have upon the numbers attending.

Summary and Conclusions

The projections above have indicated that if the trends of the past few years have been modeled realistically, the rate of growth in enrollments is likely to taper off. Nevertheless, it is quite likely that the absolute increases in the number of students will be as large in the next eight years as they have been in the past eight.

Description of the Enrollment Model for 1960-1976

The model used to allocate past enrollments for the period 1960-67 by ability and income quartiles and to project enrollments for the years 1968-76 was built up by: (1) using projections of high school graduates in each year, (2) allocating them to four ability and four socioeconomic quartiles, (3) applying estimated probabilities to the entry into the post-secondary system from each of 16 cells in the year of graduation and during subsequent years, and (4) differential survival rates from year to year were then applied to the enrollees. The total enrollment estimated by the model is a summing of these calculations:

$$1. \quad E_t = \sum_{snij}\sum\sum\sum d_{s,i,j} \cdot f_{s,i,j} \cdot r_{s,i,j} \cdot P_{s,n,i,j}\; G_{s,(t-n+1),i,j}$$

The sources used for the estimates were: (1) NCES estimates of high school graduates, (2) Project Talent data from the one-year and five-year follow-ups of the high school class of 1960, (3) U.S. Bureau of the Census attendance by age information for the period 1964-66, and (4) information about college-going intentions from two surveys conducted by Talent in the one-year follow-up survey.

TABLE 4

TOTAL, UNDERGRADUATE, AND GRADUATE DEGREE CREDIT STUDENTS – BY INCOME QUARTILE AND YEAR

(thousands of students)

Income quartile:	Low			2nd			3rd			4th			All Students		
	Total	Under-graduate	Graduate	Total	Under-graduate	Graduate	Total	Under-graduate	Graduate	Total	Under-graduate	Graduate	Total	Under-graduate	Graduate
1960	419.0	362.5	56.4	624.9	518.6	106.3	1011.5	887.2	124.2	1530.0	1313.3	216.3	3585.0	3081.6	503.4
1961	474.3	412.2	62.1	687.0	570.0	117.1	1100.4	963.4	137.0	1645.3	1407.0	239.0	3906.6	3351.8	554.8
1962	534.4	466.8	67.6	746.3	618.7	127.6	1181.2	1032.0	149.2	1747.3	1487.3	260.0	4209.3	3604.8	604.4
1963	599.1	525.9	73.2	804.1	6666.2	138.0	1257.9	1096.6	161.4	1841.3	1460.1	281.1	4502.4	3848.8	653.6
1964	680.3	599.9	80.5	882.8	730.9	151.9	1365.3	1186.6	177.7	1970.1	1660.5	309.7	4897.6	4177.8	719.7
1965	785.5	695.2	90.3	988.5	817.9	170.6	1506.4	1306.7	199.7	2143.7	1795.7	348.1	5424.1	4615.4	808.8
1966	900.0	799.3	100.6	1095.9	905.8	190.1	1646.3	1423.7	222.6	2315.7	1927.7	387.9	5957.8	5056.6	901.2
1967	1021.8	910.2	111.6	1204.8	994.1	210.7	1878.5	1540.8	246.7	2495.3	2065.3	430.0	6509.4	5510.4	999.0
1968	1137.8	1016.0	121.7	1304.6	1074.8	229.8	1911.6	1642.7	269.0	2654.3	2185.6	468.7	7008.3	5919.2	1089.2
1969	1240.7	1110.6	130.2	1391.3	1145.7	245.7	2012.8	1725.2	287.5	2775.9	2274.9	501.0	7420.7	6256.4	1164.4
1970	1339.4	1200.9	138.5	1476.1	1214.7	261.4	2116.8	1810.9	305.9	2893.7	2360.7	533.0	7825.9	6587.2	1238.7
1971	1442.5	1293.6	148.9	1569.4	1288.6	280.8	2248,7	1920.2	328.5	3052.2	2480.0	572.2	8312.8	6982.4	1330.4
1972	1547.4	1386.8	160.6	1666.3	1363.7	302.6	2395.2	2041.6	353.6	3236.4	2620.5	615.9	8845.3	7412.6	1432.7
1973	1638.6	1468.5	170.2	1750.3	1429.7	320.6	2517.5	2143.0	374.5	3383.4	2731.1	652.3	9289.8	7772.3	1517.5
1974	1718.9	1540.7	178.1	1824.4	1488.8	335.6	2621.2	2229.1	392.1	2501.4	2818.5	682.9	9665.7	8077.0	1588.7
1975	1791.5	1606.1	185.4	1891.4	1542.0	349.4	2714.6	2306.4	408.2	3611.1	2900.0	711.0	10008.5	8354.5	1654.0
1976	1858.5	1665.8	192.7	1952.6	1589.7	362.9	2805.4	2381.4	424.0	3721.6	2983.1	738.4	10338.0	8620.0	1718.0

Source: OPPE model

Estimates of the Number of High School Graduates

The number of high school graduates, separately for males and females, was taken from National Center for Educational Statistics data and estimates from the years from 1949 to 1976, published in U.S. Department of Health, Education, and Welfare, U.S. Office of Education, *Projections of Educational Statistics, 1976-77*, (Government Printing Office, Washington, D.C.: 1968), Table 17.

Estimates of High School Graduates by Ability and Income Quartile

The distribution of high school graduates by income and ability quartiles was established by splicing information from Project Talent one-year follow-up data with U.S. Bureau of the Census information.

Adjustments to Project Talent Data

Data from Project Talent were classified by grouping them four groups according to the socioeconomic index and another four groups based on aptitude test scores. Table 7 shows the tabulation of March 1960 tenth-grade males taken from the December 1963 reinterview and classified by these two variables.

TABLE 7

PERCENTAGE OF TENTH-GRADE MALES BY APTITUDE AND SOCIOECONOMIC INDEXES: MARCH, 1960

Socioeconomic Index Percentile	Aptitude Index Percentile 0-25.2	25.3-46.2	46.3-71.3	71.4-100.0	Total
0-31.5	13.9	8.2	6.0	3.4	31.5
31.6-58.61	6.3	6.3	7.9	6.6	27.1
58.7-81.4	3.5	4.4	6.9	8.0	22.8
81.6-100.0	1.5	2.1	4.3	10.7	18.6
Total	25.2	21.0	25.1	28.7	100.0

Two kinds of adjustments were made before the data were used for the enrollment model: (1) the data were adjusted to represent more precisely characteristics of students quartile by quartile, and (2) occasional statistical anomalies were smoothed.

Rates of enrollment and rates of attainment were obtained directly from Project Talent tabulations at the cumulative percentile points on the socioeconomic and aptitude indexes which were near the desired points. Langrangian interpolation procedures were then used to estimate the corresponding ratios for the desired boundaries by quartile for each of the two classifying variables independently.

TABLE 8

APPROXIMATE FAMILY INCOME DISTRIBUTION OF 1960 HIGH SCHOOL GRADUATES

Family Income Class	Percentile	Percent of 1960 High School Graduates
Under $3,337	0- 25.0	22.4
$3,338-5,625	25.1- 50.0	24.9
$5,626-8,378	50.1- 75.0	26.7
$8,398 and over	75.1-100.0	26.0
		100.0

First-year enrollments for 1960 in the year following graduation were calculated from Project Talent. This distribution by the 16 income-aptitude cells for each sex appears in Tables 9 and 10.

For previous and subsequent years to 1960, the first-time enrollment rate was adjusted on the basis of information described in Section 2 which indicated that the propensity to enroll in college had increased roughly proportionately for all income groups for the period 1939 to 1960 and had grown at different rates between 1960 and 1966. A further assumption was made that the pattern of change between 1960 and 1966 would continue for ten years to 1976; namely, that by 1973 enrollments in the first, second, and third quartile would reach the levels experienced in the next higher quartile. It was assumed that enrollment propensities in the fourth quartile would asymptotically reach 80 percent; thus the gap between current enrollments and 100 percent enrollment would be closed at the rate observed between 1960 and 1966.

TABLE 9

FIRST-TIME ENROLLMENT OF HIGH SCHOOL GRADUATES IN COLLEGE IN THE YEAR FOLLOWING GRADUATION

MALE

Income	Low	2	3	High
Low	.078	.168	.333	.536
2	.142	.245	.399	.698
3	.163	.366	.514	.753
High	.209	.362	.640	.793

TABLE 10

FEMALE

Income	Low	2	3	High
Low	.085	.125	.219	.429
2	.095	.157	.283	.565
3	.165	.249	.451	.639
High	.269	.447	.630	.854

Mathematically, this was represented as follows: upper and lower asymtotes K_1 and K_2 were chosen for each income quartile—the upper asymtote K_2 to represent the limits of growth indicated above and the lower asymtote K_1 to represent the proportions of first-time enrollment in 1944. The values of K_1 and K_2 were chosen as follows:

TABLE 11

RATE OF CHANGE IN THE PROPENSITY TO ENROLL IN COLLEGE (1944 to 1976)

Income Quartile

Year of H.S. Graduation	Low	2	3	High
1944	.160	.321	.415	.572
1945	.161	.321	.416	.574
1946	.161	.321	.417	.577
1947	.161	.322	.418	.580
1948	.162	.322	.420	.584
1949	.162	.323	.423	.588
1950	.163	.324	.426	.592
1951	.165	.326	.429	.596
1952	.167	.328	.433	.601
1953	.169	.330	.438	.607
1954	.173	.334	.444	.612
1955	.177	.339	.451	.619

(TABLE 11 (Continued)

Year of H.S. Graduation	Low	2	3	High
1956	.184	.345	.459	.625
1957	.193	.353	.469	.632
1958	.204	.363	.480	.640
1959	.219	.376	.493	.648
1960	.239	.392	.506	.656
1961	.263	.412	.522	.664
1962	.293	.435	.539	.673
1963	.327	.461	.557	.682
1964	.365	.490	.576	.691
1965	.405	.520	.595	.700
1966	.445	.550	.614	.709
1967	.483	.579	.633	.718
1968	.517	.605	.651	.727
1969	.547	.628	.668	.736
1970	.571	.647	.684	.744
1971	.591	.664	.698	.752
1972	.606	.677	.710	.760
1973	.617	.687	.721	.768
1974	.626	.695	.730	.775
1975	.633	.701	.739	.781
1976	.637	.706	.746	.788

In application of these factors, the enrollment was never allowed to exceed the ever-enrolled rate in any quartile.

TABLE 12

Income Quartile	$K_{i,1}$	$K_{i,2}$
Low	.16	.65
2	.32	.72
3	.41	.75
High	.57	.80

A logistic curve of the form

$$2. \quad y_{i,t} = K_{i,1} + \frac{C_i}{1+\ell^{a+bx}}$$

x=year and y=propensity to enroll was fitted to a pair of known points of each income quartile. The pairs of points described in Section 2 are given below:

TABLE 13

Income Quartile	1960	1966
Low	.23	.46
2	.40	.52
3	.52	.65
High	.68	.74

First-time enrollments for each year from 1960 to 1967 were weighted in the model by the ratio of

$$3. \quad r_{t,i} = \frac{y_{t,i}}{y_{1960,i}}$$

Enrollments of students in years other than the year following graduation was estimated by using the ratios published by the U.S. Bureau of the Census, the proportion of high school graduates who enrolled in college one year after graduation with those aspiring to enroll. These ratios were used in preference to Project Talent ratios of ever-enrolled because of anomalies in the Project Talent data probably due to inflation problems in the five-year follow up.

TABLE 14

RATE OF FIRST-TIME ENROLLMENT IN COLLEGE BY YEARS AFTER HIGH SCHOOL GRADUATION

(Census)

Income Quartile	1	2	3	4	5	6	7
Low	.5	.21	.13	.06	.05	.04	.02
2	.7	.13	.07	.04	.03	.02	.01
3	.8	.08	.05	.02	.02	.02	.01
High	.9	.04	.02	.02	.01	.01	----

Persistence Rates

The number of years a student was likely to be enrolled in the system was calculated by using Project Talent data and adjusting marginal totals to U.S. Bureau of the Census observed enrollment rates.

- Given enrollment by year, the maximum possible achievement for the 1960 cohort was estimated.
- The actual achievement was then calculated by using Project Talent four-year follow-up interviews. (Table 15)
- The ratio between possible and actual achievement was then used to calculate the survival rates. This is shown in Table 16.
- Beyond five years, the survival rates were projected on the basis of a straight-line fit, to a maximum of ten years or until the value reached zero.

The calculated values were then used to simulate an enrollment cohort for the year 1965. Adjustments were made in the calculated cohort survival rates to have the cohort tally estimates made by the U.S. Bureau of the Census for 1964-1966.

The new survival rates were then used to estimate cohort enrollment rates by sex, income, and aptitude for the period 1960 through 1976. These are reproduced in Table 16.

Mathematical Representation of the Model

The mathematics of calculating the cohorts is as follows:

If the symbol $a_{i,m}$ represents the number of first-time enrolled in college in the i-th income quartile, m-years after high school graduation, then from the values given in Table 14 the maximum possible proportion of the total years of attainment in K years for any given cohort of high school graduate is given by:

$$4. \quad R'_{i,k} = \frac{\sum_{M=K}^{n} A_{i,m}}{\sum_{m=1}^{n} A_{i,m}} \quad \text{for } K = 1,2 \ldots\ldots n$$

TABLE 15

RATES OF ACTUAL ATTAINMENT IN COLLEGE
GIVEN EVER ENROLLMENT
IN FIVE YEARS

Income Quartile	Aptitude Quartile	MALE Fresh-man	MALE Sopho-more	MALE Junior	MALE Senior or BA	MALE Grad. or Prof. Work	FEMALE Fresh-man	FEMALE Sopho-more	FEMALE Junior	FEMALE Senior or BA	FEMALE Grad. or Prof. Work
	1	.198	.484	.099	.203	.017	.417	.143	.163	.267	.011
1	2	.314	.229	.091	.340	.027	.244	.174	.093	.469	.020
	3	.264	.091	.066	.335	.244	.154	.211	.050	.496	.090
	4	.044	.188	.103	.403	.263	.073	.127	.079	.586	.135
	1	.272	.253	.148	.310	.016	.179	.522	.074	.220	.006
	2	.300	.255	.140	.216	.089	.134	.273	.190	.295	.108
2	3	.131	.163	.168	.401	.137	.279	.130	.155	.368	.068
	4	.145	.076	.124	.428	.227	.127	.109	.168	.398	.199
	1	.489	.146	.198	.141	.026	.422	.039	.130	.322	.087
	2	.223	.244	.132	.336	.065	.187	.349	.098	.320	.046
3	3	.194	.172	.086	.322	.226	.199	.151	.137	.425	.089
	4	.050	.136	.117	.396	.300	.063	.154	.119	.428	.236
	1	.201	.164	.158	.369	.108	.050	.496	.265	.144	.045
	2	.137	.199	.132	.333	.199	.213	.230	.079	.277	.202
4	3	.050	.207	.120	.409	.215	.180	.146	.141	.422	.112
	4	.040	.077	.106	.332	.446	.025	.107	.065	.567	.238

Source: Project Talent 5th Year Interview (unpublished).

Table 16

PERSISTENCE RATES IN COLLEGE IN FIRST FIVE YEARS AFTER
HIGH SCHOOL GRADUATION BY SEX, INCOME, AND APTITUDE

Income Quartile	Aptitude Quartile	MALE 1	MALE 2	MALE 3	MALE 4	MALE 5	FEMALE 1	FEMALE 2	FEMALE 3	FEMALE 4	FEMALE 5
	1	1.0	.530	.360	.290	.030	1.0	.598	.498	.371	.021
	2	1.0	.720	.520	.360	.050	1.0	.798	.658	.655	.038
1	3	1.0	.770	.750	.580	.120	1.0	.894	.719	.784	.171
	4	1.0	.850	.800	.700	.210	1.0	.978	.904	.965	.257
	1	1.0	.700	.520	.380	.020	1.0	.848	.323	.264	.008
	2	1.0	.730	.480	.360	.120	1.0	.894	.639	.471	.150
2	3	1.0	.900	.770	.630	.190	1.0	.744	.637	.509	.094
	4	1.0	.890	.850	.770	.310	1.0	.902	.824	.697	.267
	1	1.0	.520	.380	.180	.030	1.0	.590	.562	.451	.105
	2	1.0	.790	.550	.440	.080	1.0	.830	.484	.404	.056
3	3	1.0	.820	.660	.600	.270	1.0	.819	.679	.567	.108
	4	1.0	.970	.800	.760	.360	1.0	.957	.816	.732	.286
	1	1.0	.800	.660	.530	.120	1.0	.951	.468	.199	.050
	2	1.0	.860	.680	.560	.220	1.0	.788	.574	.505	.222
4	3	1.0	.950	.770	.660	.240	1.0	.822	.696	.563	.123
	4	1.0	.960	.910	.820	.500	1.0	.976	.896	.847	.261

TABLE 17

ADJUSTMENT FACTORS ON ENROLLMENT RATES

Years After High School Graduation

	1	2	3	4	5	6	7	8*	9*	10*
Males	1.0	1.0	1.1	1.12	1.25	1.25	1.5	8.25	8.25	8.25
Females	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0425	1.0425	1.0425

*These years in the Census data represent five-year age intervals, hence, the larger factor.

Similarly, using the values in Table 15 the actual attainment in K years of enrollment of the cohort of high school graduates may be calculated by sex, income, and aptitude quartiles.

$$5. \quad R''_{s,i,j,k} = \frac{\sum_{m=k}^{n} b_{s,i,j,m}}{\sum_{m=1}^{n} b_{s,i,j,m}} \qquad \begin{array}{l} \text{for } K = 1,2....m \\ \text{for } S = 2,2 \end{array}$$

The survival rates in Table 16 for each of the 16 income and aptitude quartiles will then be given by the ratio:

$$6. \quad R_{s,i,j,k} = \frac{R''_{s,i,j,k}}{R'_{i,k}}$$

The 16 rates of enrollment in college over K years, given the probability of enrollment at any time after graduation was calculated from the above by "entering" students for the first time m years after high school graduation and applying the survival ratio $R_{s,i,j,k}$ to those enrolled and then summing for each of the K years. The equations giving the 16 enrollment curves are derived as follows:

Let the matrix

$$7. \quad C_{s,i,j,m,k} = A_{i,m} R_{s,i,j,(K-m+1)}$$

where $a_{i,m}$ are the entries in Table 14 and the R_s are the survival curves calculated by equation 6

Summing the resulting C matrix for each of the K years over m yields the enrollment rates for each of the 32 sex, income-aptitude cells.

$$8. \quad S_{s,i,j,k} = \sum_{m=1}^{n} C_{s,i,j,m,k} \qquad \text{for } K=1,2....n$$

Since, however, the Talent distribution of first-time enrollment by income and aptitude appeared to be more reliable than the "ever-enrolled" data obtained from the five-year survey, the curve was adjusted to apply to first-time enrollment by transformation:

$$9. \quad P_{s,i,j,k} = \frac{S_{s,i,j,k}}{S_{s,i,j,1}}$$

The Talent data accounted for only the first five years of enrollment after high school graduation, the enrollment curve was therefore adjusted for years beyond five to simulate the total population included in the U.S. Bureau of the Census estimates. The adjustment factors for males and females are given in Table 17.

The enrollment rates, based on first-time enrollment in the first year following high school graduation, for males and females, is presented in Table 18.

Full-time equivalent enrollment students were estimated from total enrollment by weighting the estimated number full-time students by the ratio of years attained to years attended, obtained from Project Talent data. The ratios are given in Table 19.

Estimates of Full-Time Students

Full-time enrollment by ability and income quartile was calculated from total enrolled by applying differential rates of full-time attendance for each of the 16 groups. It was assumed that part-time students carried a one-third load. Given the above ratios of years attained per years attended, it was possible to calculate the full-time students with the equation:

$$10. \quad y_{s,i,j} = a_{s,i,j} + \frac{1}{3}(1.0 - a_{s,i,j})$$

where:

y = Years attained per years enrolled.
a = Ratio of full-time enrolled to total enrolled.

These estimates were applies to the 1964 cohort, the year for which NCES did their last full-time/part-time census and were scaled to reproduce the NCES estimates. The scaling factor used was .92 on the years attained to years attended ratio.

The estimates used in the model appear in Table 20.

Graduate and First-Professional Degree Students

Estimates of graduate and first professional students are derived in the model in two steps. For the years 1960-1967, NCES estimates of total graduate students by sex were adopted. For the period 1968 through 1976, they are projected as a function of total enrollment based on the fitting of the following function to the period 1960 through 1967:

$$11. \quad y_{s,t} = as^{b}_{s,t}$$

Where $s_{s,t}$=total enrollment and $y_{s,t}$=NCES estimate of graduate students, the calculated coefficients and exponents were:

a=.054 for males, .027 for females.

b=1.15 for males, 1.18 for females.

The estimated number of graduate students was increased 25 percent, to account for first professional students, in line with observed differences between students enrolled in programs in the fifth year of college as estimated by the U.S. Bureau of the Census and the total graduate enrollment reported by NCES for the period 1964-1966.

The number of full-time graduate students by income quartile and by year was estimated by weighting the relative distribution of all full-time, total enrolled graduate students by an appropriate ratio in each income quartile.

In 1965, the ratio of full-time to total graduate students was 44 percent. The ratio of all full-time to total enrolled students was 63 percent. Hence, a weight of .44=.7 was applied to the ratio of full-time to part-time students in each income cell by sex. For instance, in 1965, the estimate of all full-time students to total male students in the high income cell was .95. It was, hence, estimated that the proportion of full-time graduate students was .95 x .7 = .665 of total enrolled graduate students.

The estimated number of full-time graduate students was subtracted from the estimated number of total full-time students to derive the number of full-time undergraduate students enrolled.

TABLE 18

ENROLLMENT RATES IN COLLEGE BASED ON FIRST-TIME ENROLLMENT

Income Quartile	Aptitude Quartile	1	2	3	4	5	6	7	8	9	10
						MALE					
1	1	1.00	.95	.84	.70	.41	.26	.16	.08	.04	.01
	2	1.00	1.14	1.08	.89	.52	.33	.21	.11	.05	.02
	3	1.00	1.19	1.33	1.21	.75	.51	.30	.18	.10	.04
	4	1.00	1.27	1.42	1.38	.91	.72	.43	.25	.14	.07
2	1	1.00	.89	.75	.60	.23	.13	.08	.04	.02	.01
	2	1.00	.92	.72	.58	.32	.15	.09	.05	.02	.01
	3	1.00	1.09	1.04	.92	.48	.34	.15	.09	.04	.02
	4	1.00	1.08	1.12	1.07	.63	.56	.37	.16	.08	.04
3	1	1.00	.62	.49	.28	.22	.12	.11	.02	.03	.01
	2	1.00	.89	.69	.57	.41	.30	.18	.11	.06	.02
	3	1.00	.92	.80	.74	.42	.31	.29	.18	.10	.05
	4	1.00	1.07	.96	.93	.54	.51	.47	.22	.12	.06
4	1	1.00	.84	.72	.57	.37	.15	.09	.05	.02	.00
	2	1.00	.90	.74	.63	.58	.34	.10	.06	.03	.01
	3	1.00	.99	.83	.74	.64	.49	.27	.08	.04	.01
	4	1.00	1.00	.97	.90	.59	.53	.45	.33	.21	.07
						FEMALE					
1	1	1.00	1.02	1.01	.86	.48	.30	.19	.10	.05	.02
	2	1.00	1.22	1.25	1.26	.68	.43	.26	.15	.08	.03
	3	1.00	1.31	1.35	1.44	.89	.53	.32	.19	.11	.05
	4	1.00	1.40	1.57	1.72	1.11	.65	.39	.24	.14	.06
2	1	1.00	1.03	.58	.47	.18	.11	.07	.03	.01	.00
	2	1.00	1.08	.91	.74	.40	.18	.11	.06	.03	.01
	3	1.00	.93	.88	.76	.34	.17	.10	.06	.03	.01
	4	1.00	1.09	1.09	1.00	.58	.24	.14	.08	.04	.02
3	1	1.00	.69	.68	.57	.23	.09	.06	.04	.02	.01
	2	1.00	.93	.63	.53	.17	.09	.06	.03	.02	.01
	3	1.00	.92	.82	.71	.25	.11	.07	.04	.03	.01
	4	1.00	1.06	.97	.90	.46	.14	.09	.06	.04	.02
4	1	1.00	1.00	.53	.26	.10	.04	.02	.01	.00	.00
	2	1.00	.83	.63	.57	.29	.05	.03	.02	.01	.00
	3	1.00	.87	.75	.63	.19	.05	.03	.02	.01	.00
	4	1.00	1.02	.96	.93	.35	.07	.05	.03	.01	.00

TABLE 19

RATIO OF YEARS ATTAINED PER YEARS ATTENDED

	MALE				FEMALE			
		Aptitude				Aptitude		
Income Quartile	Low	2	3	High	Low	2	3	High
Low	.691	.711	.774	.886	.726	.778	.739	.849
2	.676	.744	.819	.901	.738	.763	.785	.889
3	.679	.767	.844	.924	.714	.758	.835	.897
High	.661	.831	.894	.946	.763	.816	.867	.950

TABLE 20

RATIOS OF FULL-TIME ENROLLMENT TO TOTAL ENROLLMENT BY SEX, INCOME, AND APTITUDE

MALE

Income Quartile	Aptitude Low	2	3	High
Low	.454	.481	.569	.723
2	.433	.526	.630	.743
3	.437	.559	.665	.775
High	.412	.647	.733	.805

FEMALE

Income Quartile	Aptitude Low	2	3	High
Low	.502	.574	.520	.672
2	.519	.553	.583	.727
3	.486	.546	.652	.738
High	.553	.627	.696	.810

TABLE 21

MEAN STUDENT COSTS BY SEX AND INCOME QUARTILE

Income Quartile	MALE Tuition Expense	MALE Living Cost	FEMALE Tuition Expense	FEMALE Living Cost
Low	319	559	290	509
2	411	728	394	762
3	333	557	311	528
High	401	772	397	762
Annual Increment	r_T=.06	r_L=.02	r_T=.06	r_L=.02

FINANCIAL ASSISTANCE REQUIREMENTS

The total need for assistance in the financing of the cost of higher education for full-time undergraduate students, the difference between total student costs and total family contribution was estimated on three alternative bases in the model for Low and High enrollment projections.

Needs based on mean expenditures. Mean costs reported by full-time students for Project Talent by sex were calculated for each of the four income groups, and Mean parental contributions were calculated for each income quartile from the College Scholarship Service expected of parental support for college expenses for families with two children. [3]

The estimated college costs from Talent were in 1960 dollars, and these were incremented at an annual rate of 6 percent for tuition expenses and at 2 percent for living costs. Parental incomes were allowed to grow at 5 percent annually, assuming a 2 percent rate of inflation. Parental contributions, in 1966 dollars, were incremented at a varying annual ratio as shown in Table 22 in such a way that the 1966 proportion of the real income was actually set aside for college expenses. The gap, $g_{t,i}$ was calculated as follows:

Let

$$T_{t,s,i} = \sum_j {}^{T}C_{s,i} \cdot E_{s,t,i,j} \cdot (1+n_t)^t$$

$$L_{t,s,i} = \sum_j {}^{L}C_{s,i} \cdot E_{s,t,i,j} \cdot (1+r_L)^t$$

$$K_{t,s,i} = \sum_j P_i \cdot \sum_s E_{s,t,i,j} \cdot (1+r_k)^t$$

Then

12. $$g_{t,i} = \sum_s T_{t,s,i} + L_{t,s,i} - K_{t,s,i}$$

If T + L - K was < 0, then $g_{t,i}$ was made = 0.

T = Tuition cost by sex and income quartile.
L = Living cost.
K = Parental contribution by year, second income quartile.
E = Estimated enrollment by year, sex, income, and aptitude.
C = Student costs for tuition or living expenses as indicated in the superscript.
P = Parental contribution.

TABLE 22

PARENTAL CONTRIBUTIONS TO COLLEGE EXPENSES BY INCOME QUARTILE

	INCOME QUARTILE Low	2	3	High
r_k =	.22	.07	.08	.014
1966	$ 20	$ 750	$1,360	$3,500
1967	25	795	1,458	3,565
1968	35	843	1,633	3,630
1969	43	898	1,730	3,698
1970	56	977	1,887	3,767
1971	68	1,071	2,023	3,836
1972	82	1,158	2,134	3,908
1973	96	1,297	2,281	3,980
1974	111	1,319	2,464	4,054
1975	120	1,410	2,751	4,129
1976	149	1,532	2,869	4,205

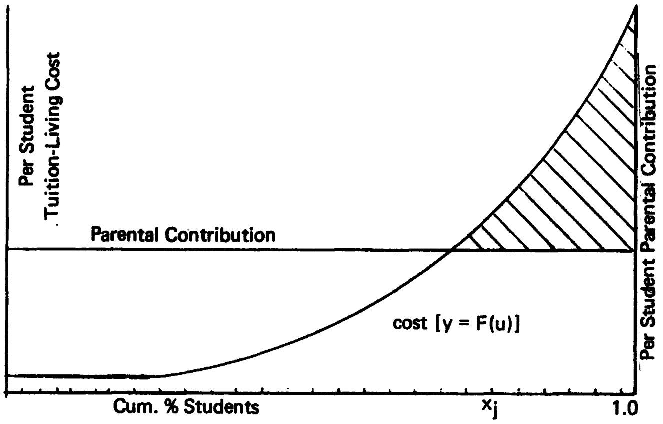
2. Distribution of college-enrolled male and female students by tuition cost and living expenses were calculated from Project Talent data. These distributions are given in Tables 23 and 24. On the assumption that students who pay the lowest tuition also have the lowest living cost, 16 combined tuition-living cost curves were calculated for men and women by quartile.

An exponential curve of the form:

13. $$y = ae^{bx}$$

(where x = Percentile scale; y = Combined cost per student) was fitted to each of the 16 cost curves, and the "GAP" was then recalculated on the following basis:

Given the two cumulative curves on the graph below, representing the combined student costs and the parental contribution, the shaded area represents the total gap between student costs and parental contribution.



The "GAP" in a given income quartile is the sum of the gaps calculated for men and women in that income quartile, by the equation:

$$14. \quad g_{t,i} = E_{t,i} \cdot \sum_{s} \left(\int_{j}^{1\cdot 0} f_{s,i}(u)\, du - (1-x_j)P_i \right)$$

TABLE 23

DISTRIBUTION OF COLLEGE-ENROLLED STUDENTS BY SEX, FAMILY INCOME QUARTILE, AND TUITION EXPENSES

MALE

Tuition Expense	Income Quartile 1	2	3	4
50	.029	.024	.022	.018
75	.090	.069	.060	.039
150	.146	.124	.106	.085
250	.195	.162	.130	.117
400	.166	.197	.207	.175
625	.145	.115	.130	.130
875	.129	.142	.128	.125
1250	.082	.134	.157	.190
1500	.017	.034	.060	.123

TABLE 23 (Continued)

DISTRIBUTION OF COLLEGE-ENROLLED STUDENTS BY SEX, FAMILY INCOME QUARTILE, AND TUITION EXPENSES

FEMALE

Tuition Expense	Income Quartile 1	2	3	4
50	.032	.027	.020	.012
75	.076	.054	.060	.041
150	.141	.102	.111	.075
250	.173	.369	.154	.103
400	.216	.156	.201	.166
625	.152	.106	.126	.129
875	.092	.090	.135	.139
1250	.108	.080	.144	.202
1500	.011	.016	.049	.134

TABLE 24

DISTRIBUTION OF COLLEGE-ENROLLED STUDENTS BY SEX, FAMILY INCOME QUARTILE, AND LIVING COSTS

MALE

Living Costs	Income Quartile 1	2	3	4
300	.428	.373	.331	.216
400	.209	.162	.163	.142
625	.207	.250	.232	.195
875	.110	.143	.173	.213
1250	.041	.062	.084	.170
1750	.001	.007	.013	.044
2250	.001	.002	.003	.013
2750	.000	.001	.000	.004
3000	.000	.002	.002	.004

FEMALE

Living Costs	Income Quartile 1	2	3	4
300	.439	.383	.288	.201
400	.239	.192	.179	.147
625	.229	.245	.260	.222
875	.066	.127	.181	.221
1250	.023	.043	.074	.152
1750	.001	.007	.013	.037
2250	.003	.001	.004	.013
2750	.000	.001	.001	.002
3000	.000	.001	.001	.006

3. Student costs and the financial gap were calculated on the basis of a third set of alternative assumptions as follows:

The number of students in the lower and upper two-year levels in all institutions were calculated from ratios obtained from Project Talent data and from projections of full-time undergraduate students.

TABLE 25

FULL-TIME UNDERGRADUATE STUDENT IN ALL INSTITUTIONS IN ACADEMIC LEVEL

Income Quartile	Lower 2-Years	Upper 2-Years
Low	.70	.30
2	.65	.35
3	.60	.40
4	.55	.45

Average student costs were calculated from cost estimate given in Table A-6 of *Students and Buildings* [4] and the entries in Table 25 above by the equation.

$$15. \quad \bar{C}_{i,m} = \Sigma\, d_{i,j} \cdot C_{j,m}$$

where $\bar{c}$ = average cost, d = the entries in Table F-5 and c = the cost factors from *Students and Buildings* (see Tables 26 and 27)

TABLE 26

ASSUMED COST FACTORS

Academic Level	Tuition m = 1	Living Exp. m = 2
Lower 2-yrs.	$103	$1,000
Upper 2-yrs.	278	1,283

TABLE 27

AVERAGE STUDENT COSTS

Income Quartile	Tuition	Living Exp.
Low	$156	$1,084
2	164	1,100
3	173	1,112
4	182	1,175

The projected student costs for each year by income quartile may be calculated from the above by the equation:

$$16. \quad C_{t,i} = \sum_{m} E_t \cdot \bar{C}_{i,m} \cdot (1+rm)^t$$

and as in hypothesis one the gap, $g_{t,i}$, is calculated as the difference between student costs and parental contribution.

NOTE 1

Explanation of Symbols (Equation 1)

E = Total Enrollment.

d = Distribution of High School Graduates by Sex, Income, and Aptitude.

f = First-Time Enrollments in College in Year Following High School Graduation by Sex, Income, and Aptitude.

r = Rate of Growth in the Propensity of High School Graduates to Enroll in College by Income, 1944-1976. Table 11. (p.).

P = Enrollment Rates in College in Years after High School Graduation, Given First-Time Enrollment in First Year Following High School Graduation by Sex, Income, and Aptitude. Tables 9, 10. (p.).

G = Estimated Number of High School Graduates by Sex. Source: U.S. Office of Education Publication Projections of Educational Statistics to 1976-1977.

Subscripts:

t = Current Year.

s = Sex.

i = Income Quartile (1 = low, 4 = high).

j = Aptitude Quartile (1 = low, 4 = high).

n = Years Since High School Graduation.

REFERENCES

[1] It should be noticed that the realtionship between total and full-time equivalent enrollment varies from year to year in the projections, and is different from NCES estimates. Full-time equivalent enrollments have been imputed by the model on the basis of attendance rates by quartile. NCES projects a slightly changing mix between full-time and part-time students. See Technical Appendix C.

[2] National Center for Educational Statistics, *The Academic and Financial Status of Graduate Students,* (Washington, D.C., Government Printing Office), 1965.

[3] Financing a College Education. *A Guide for Counselors.* 1966. College Entrance Examination Board.

[4] *Students and Buildings,* U.S. Department of Health, Education, and Welfare, U.S. Office of Education, Government Printing Office, Washington, D.C., 1968.

THE IMPACT OF PARENTAL AND EDUCATIONAL INVESTMENTS UPON STUDENT ACHIEVEMENT

Dennis J. Dugan*

Parents provide a home environment that contributes to the scholastic success of their child (ren). This relationship between home-environment and achievement is well-known, and scholastic achievement has generally been assumed to flow directly from the socio-economic characteristics of parents and a complement of school resources. In this paper the mechanism that parents employ to transmit the "potential-for-learning" to their children is considered explicitly within the context of a theoretical model of the education process. Further, a method to measure the educational services that parents provide their child(ren) in the home environment is presented.

The first section of the paper is devoted to the derivation of estimates value of parental investments into the education process. The value of these investments is determined by the opportunity cost of the time that parents devote to the stimulus-response-reinforcement activities of their child(ren) during both the pre-school and formal-education years. The analysis of the components of the total cost of education, including the educational services contributed by the parents, indicates that school expenditures account for less than 50 percent of the total investment in education. In the second section of the paper, the education process is defined in terms of a theoretical model which explains achievement differences by variations in the contributed educational services of parents and school resources. The model is estimated empirically for the white and nonwhite student populations, and the relative effectiveness of educational investments is determined. The statistical results indicate that the educational services of parents are highly significant determinants of scholastic achievement. The empirical results also demonstrate that nonwhite scholastic performance would be significantly increased if the nonwhites had the same level of parental services and school resources as the white students have.

TOTAL INVESTMENT IN THE EDUCATION PROCESS

While the significance of genetically or congenitally determined resources which the individual child has at his disposal is not discounted herein, the main concern of this paper is with the educational processes of the home environment--factors particularly relevant in determining the form his ability will take. The education process is conceived to be a complex process of long gestation in which school and parental investments are combined to produce scholastic achievement, the final product of the educational system.

Taking a close look at the entire educational process, one sees a long formation period in which the student accumulates educational investments and builds upon his previous experiences. This experience is not limited to formal schooling alone and, undoubtedly, it relies heavily upon the home environment of the child and his pre-dispositions towards learning that are acquired in the home. If, in fact, education is more than time and resources provided by the school, then current expenditures per student do not reflect all, nor necessarily a large portion, of the costs that are directly relevant to the educational process. Therefore, "equal educational opportunity" is not brought about simply by equalizing school expenditures among students.

There is substantial evidence[1] that pre-school experience and home environment contribute to a child's success in school. These two activities are dependent upon services rendered by the child's mother and father--services that require time and effort that could be devoted to alternative activities. The value of such services and their marginal contribution to educational output are particularly important for public educational policy which has the goal of presenting "equal educational opportunity to all members of our society."

Under this expanded view of the educational process, total educational investment may be substantially greater than formal schooling expenditures. This is especially true if the mother spends five to six years--which she ordinarily does--preparing the child for formal schooling. The value of the educational services provided by the mother is a social cost which would be reflected by a higher level of Gross National Product (GNP) had the mother been working in an occupation that is tallied in national income accounts. A mother has alternative employment opportunities that are consciously by-passed to stay with her child(ren). The net value of the mother's educational services to society is not zero, and they may be approximated by determining the value of her services in an alternative form of employment,[2] that is, by calculating the earnings that she foregoes to provide the educational services for her child(ren).

The father also provides educational services to the child in the home. These paternal services have a social value and that value is estimated by the father's opportunity cost of extended participation in the labor force.

The total educational investment in a student, therefore, is the accumulated sums of the educational contributions of the mother, the father, and the school. The total investments of these three factors are significant not only because they sum to the aggregate educational investment, but also because they indicate the nature of the separate resources that may have an impact upon scholastic achievement.

Measures of Parental Investment

Age and education are used as predictors of the foregone earnings of parents, and these earnings are used as a measure of parental opportunity costs. Age-earnings profiles by level of education in the U.S. typically resemble those in Figure 1.

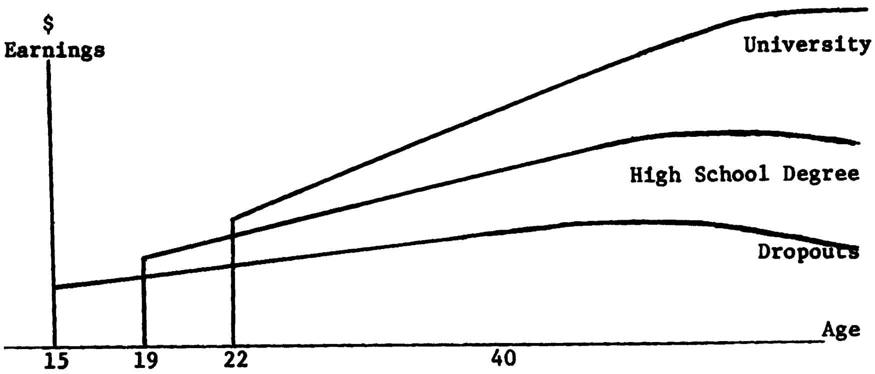


Figure 1. Typical Age-Earnings Profiles by Levels of Education

These profiles have three striking characteristics:

(1) Earnings, irrespective of the level of education, increase with age up to a maximum and then decline. The obvious explanation is that age acts as a proxy for work experience, raising earnings until infirmity or educational obsolescence sets in.

(2) The higher the educational attainment, the higher the starting salary and the steeper the rise in earnings throughout the early phases of working life.

(3) The higher the educational attainment, the later the year at which maximum earnings are reached and the gentler the decline of earnings from the maximum point.

For the mother who provides educational services in the pre-school years of her children, the value of her contributed services may not be the entire amount of income that she is capable of earning by offering her services to the "open" labor market. The mother provides other services in the home besides those that are directly attributable to the education formation of her offspring. Household activities such as washing, cooking, and cleaning are productive economic activities that should not be included in the "opportunity cost" of the maternal contributed educational services. A study of the time-allocation of housewives with young children, under age six, indicates that the mother spends approximately 43 percent of her time in activities directly related to her children--time which is devoted to some type of communication or contact between herself and her child(ren).[3] These stimulus-response-reinforcement activities in the home provide a potential-for-learning to the child and may be an important determinant of achievement in school.

A recent study[4] indicates that when children are in school the mother devotes 5 percent of her economically productive time (the time she would have spent in the labor force if she worked full time) to the educational activities of her children. It was assumed that the father offers a constant 5 percent of his working time to the education activities of his children during their pre-school and formal education years. These parental services are divided up among the children in the family.

The worth of the contributed services of the parents for a given period of time $\underline{t}$ is earnings from alternative employment opportunities, that is, the "opportunity cost" of participating in the labor force more fully instead of devoting these resources to the child. The opportunity cost concept need not consider more participation in the labor force as the only viable alternative to homework. Leisure time, with its implicit value, is another alternative use of time that may be pursued by parents.[5] The Contributed Educational Services of the Mother is defined as:

$$(1) \qquad M_t = a_t Y_t$$

where $\underline{Y_t}$ is the mother's "opportunity cost," the income she could be earning by participating fully in the labor force and $\underline{a_t}$ represents a combination of factors that determine how much of the mother's time is devoted to each of her children. These factors include: (1) the proportion of a mother's time devoted to educationally-related activities (as opposed to household chores), and (2) the number of children among whom the mother's time is divided. The $\underline{a_t}$ is not a constant, and it will vary over time when a mother has more children in the home and when these children leave the pre-school environment.

The mother's contributed educational services are divided into two distinct units: the pre-school portion in which the mother spends a considerable amount of time with her pre-school children; and the mother's time provided during the formal schooling years of her children. The contributed services of the mother may now be looked upon as the sum of these two time intervals, one from birth to age six and the other from age six to age eighteen.

$$(2)\quad M = \sum_{t=0}^{5} a_t Y_t (1+i)^{18-t} + \sum_{6}^{18} b_t Y_t (1+i)^{18-t}$$

where $\underline{a_t}$ represents the fraction of the mother's time devoted to a child in the pre-school situation and $\underline{b_t}$ represents the fraction of the mother's time devoted in the formal schooling period. For the pre-school period, it is assumed that the mother spreads 43 percent of her productive time among the pre-school children in the home. Vital statistics from the U.S. Census of Population gives a breakdown of average spacing of children and the number of children by the educational attainment and race of the mother. That information determines the size of the constant, $\underline{a_t}$.[6] $\underline{b_t}$ is a fraction that indicates the proportion of the mother's time devoted to her children during their school years. It is assumed that mothers divide 5 percent of their time equally among their school-age children.

The contributed educational services of parents are accumulated at a positive rate of interest, $\underline{i}$, to indicate that alternative investment opportunities are available to parents and that these parental resources need not be devoted to embodying human capital in their children. For estimation purposes, $\underline{i}$ is assumed to be a constant 6 percent.

The father's contributed educational services is derived in the same manner:

$$(3)\quad F = \sum_{t=0}^{18} c_t I_t (1+i)^{18-t}$$

where $\underline{F}$ is the accumulated value of the father's contributed services, $\underline{I}$ the "opportunity cost" of the services or the father's salary, and $\underline{c_t}$ is the portion of the father's time devoted to the educational activities of each child.

The accumulated maternal contributed services to a student at different grade levels according to the educational attainment of the mother is shown in Table 1. This table is based upon 1960 Bureau of the Census data which cross-classifies the years of educational attainment and number of children of mothers. Appendix B presents these data which show the inverse relation between a mother's educational attainment and the number of her children. At grade 1, the student has accumulated varying amounts of educational services from his mother depending on her level of educational attainment. The services rendered in the pre-school years from a mother with five plus years of college are worth $9,322 according to our opportunity cost concept of valuing these services. A mother with 0-7 years of education, by contrast, has provided $2,723 worth of educational services to a first grader.

Over the education cycle of the student, the accumulated educational services of the mother grow as she provides a limited number of educational services in the home--an average of 20 minutes per day compared to an average of three hours per day in the pre-school years of her children.

There are substantial differences in the value of contributed educational services per child between whites and nonwhites. The nonwhite mother's level of educational attainment is, on the average, lower than that of the white mother and her earning power at a given level of educational attainment is generally lower than that of a white mother. She is also likely to have more children. Part of these differences between white and nonwhite earnings structures may be attributed to discrimination, but part of it is due to the poorer quality of the nonwhite's educational experience. Therefore, for this analysis of the total cost of education, the population was stratified according to white and nonwhite classifications. The average level of educational attainment of a white mother is 11.2 years and it is 8.5 years for nonwhite. At grade one the accumulated educational services of white and nonwhite mothers with the above-mentioned years of educational attainment are $4,991 and $2,508, respectively. Since the average nonwhite mother has more children than the average white mother, the differences between the embodied capital of white and nonwhite first graders are quite pronounced.

The Total Costs of Education

The mother's and father's contributed educational services and the accumulated resources expended by the school[7] are tallied in Table 2 for grade 12, by the level of parents' educational attainment. At grade 12 the accumulated per pupil school expenditure totals $6,921 for the average student. School costs, as a percent of total educational costs, vary from 51 percent for a student whose parents have 0-7 years of education to 21.8 percent in the case where both parents have attained five plus years of college. The mother's contribution to total cost ranges from 34 percent to 49 percent for her 0-7 years and five plus years of college, respectively. The father's factor share increases with his level of educational attainment, between 12 and 21 percent.

The total costs of educational services (school plus contributed educational services of parents) for all grade levels are in the order of two to three times larger for students

TABLE 1

Value[8] in 1965 of Mother's Accumulated Contributed Educational Services for Mother's Educational Attainment, by Grade of Child

	Grade			
Mother's Education	1	6	9	12
0-7 years	$2,723.50	$ 3,411.61	$ 4,126.38	$ 4,988.87
8 years	3,378.67	4,231.37	5,135.45	6,235.03
1-3 years of high school	3,972.14	5,012.26	6,093.91	7,408.66
4 years of high school	6,963,71	8,898.18	10,796.85	13,080.38
1-3 years of college	7,090.81	9,050.96	10,995.03	13,362.96
4 years of college	9,043.76	11,560.35	14,075.54	17,147.79
5+ years of college	9,321.80	11,919.85	14,643.85	17,977.84

TABLE 2

Factor Shares and Total Cost of Educational Inputs of Grade 12 for Educational Attainment of Parents (All figures are %'s, unless otherwise noted)

Education of Parents	Mother (Pre-school)	Father	Mother (Other)	School	Total Costs
0-7 years	34.30	12.17	2.49	51.04	$13,540.83
8 years	36.82	15.49	3.22	44.46	15,567.63
1-3 years of high school	32.83	18.24	9.44	39.49	17,527.01
4 years of high school	49.49	17.58	4.41	28.52	24,269.02
1-3 years of college	48.24	19.52	4.78	27.46	25,203.26
4 years of college	49.67	22.21	5.75	22.37	30,942.18
5+ years of college	49.11	21.45	7.61	21.84	31,696.98

with parents having completed college than those having only some elementary education. At Grade 1, the ratio is approximately three to one and it diminishes slightly as the student proceeds to higher grades. The disadvantaged student, one whose parents have little education, remains disadvantaged throughout his educational career, even though this disadvantage may diminish over time. This analysis of the total costs of education indicates that equalization of school resources devoted to the culturally advantaged and disadvantaged may not eliminate educational deficiencies nor insure equality of educational opportunity.

A NEW MODEL OF ACHIEVEMENT

The Equality of Educational Opportunity Survey(6), popularly known as the Coleman Report, stimulated a great deal of research concerning the determinants of scholastic achievement. While Coleman tended to minimize the influence of the characteristics of schools upon achievement, Bowles(4) and Bowles and Levin(5) emphasize the relationship between various school inputs and achievement. Specifically, Bowles(4) adjusts for the combined effects of the student's genetic ability and environmental influences prior to age six be introducing first grade verbal achievement test scores. He arrives at the flow of scholastic achievement from grade 1 to grade 12 for black male students. Bowles' empirical "education production function" links a number of school variables, such as teacher's verbal ability score and science lab facilities, to achievement.

In this section a theoretical model of the education process is presented and one of the determinants of school outcomes is the home-embodied capital of students. Since the education process builds on what has gone before, the determinants of achievement should be specified in a form that reflects this gradual enlargement of the student's education base. Therefore, the appropriate variables to explain differences in school outcomes are the accumulated educational investments of parents and accumulated school resources. Thus, the education process is envisaged here as an educational "learning-by-doing" model where parental and school investments, both past and present, act as positive reinforcement for the student's academic success.

The model may be specified formally in the following functional form:

$$(4) \quad A_i = f(M,F,S)$$

where A_i is a suitable measure of school outcomes for grade i, M and F are respectively the contributed educational services of the mother and the father, and S is the accumulated school resources provided the student. Since measures of achievement, the outcome of schooling, is a stock variable, parental and school services are expressed in stock terms to maintain consistency. The hypothesis to be tested may be stated simply: the contributed educational services of parents and school resources jointly determine the scholastic achievement of a student. This hypothesis has been postulated in such a manner that it is empirically testable under a number of alternative formulations of the basic relationship in equation (4).

The exact functional form that the relationship in equation (4) assumes is not predetermined by a theory of learning. Two mathematical forms of that relationship, however, have immediate theoretical appeal because of the interpretations that may be given to their empirical results. One functional form of the relationship between achievement and parental and school investments is linear where the variables enter separately and linearly:

$$(5) \quad A_i = \alpha_{i0} + \alpha_{i1} M + \alpha_{i2} F + \alpha_{i3} S + U_i$$

where U_i is the unobserved, statistical disturbance term and the i's refer to school grades. The coefficients of this linear model are statistically estimated by the application of multiple regression analysis. Since all the independent variables in the linear model are scaled in dollar terms, their estimated coefficients will yield a direct answer as to what scarce educational resource yields the largest increment in achievement. The α's may appropriately be called "benefit-cost" coefficients because each indicates the change in school outcomes due to a change in the unit cost of investment in that resource. Underlying this linear model is the assumption that the school's influence upon achievement is independent of the home environment and the home-environment influence is independent of the school. Insofar as the linear model truly represents the education process, then the variable with the largest α would merit all available educational resources to achieve the greatest impact upon school outcomes.

Another function form that this relationship may assume is the following nonlinear one:

$$(6) \quad A_i = \beta_{i0} M^{\beta i1} F^{\beta i2} S^{\beta i3} U_i$$

which is easily transformed into a linear-in-the-logarithms function which may be empirically estimated by regression analysis:

$$(7) \quad \ln A_i = \ln\beta_{i0} + \beta_{i1}\ln M + \beta_{i2}\ln F + \beta_{i3}\ln S + V_i$$

where $\ln$ refers to the natural logarithm of the appropriate variable, and V_i is the natural logarithm of U_i.

This logarithmic model has theoretical appeal because it takes into account diminishing returns to the resources, that is, the intensity of an educational service, provided by a parent or the school, may be instrumental in determining its marginal contribution to scholastic achievement. This model also allows for some interaction among the independent variables in determining school outcomes and the degree of substitution among the determinants is restricted by technological and structural considerations. In the linear model the degree of substitution among those who provide educational services is virtually unlimited.

The alternative formulations of the learning model that appears in equations (5) and (7) are theoretically consistent with the hypothesis stated above. Both take into account the previous educational experiences of the student and build a scholastic record upon these prior investments. Each formulation of the model has its specific theoretical interpretation, and the statistical analyses of the two equations may indicate what model best explains the hypothesis.

This paper utilizes the information from the Equality of Educational Opportunity Survey to identify the home characteristics of students and to associate these with the results of achievement tests given to the students. This survey of the Nation's public elementary and secondary schools was conducted in the Fall of 1965 by the U.S. Office of Education, at the direction of Congress. The survey entailed the testing and surveying of about 650,000 students in some 4,000 public schools throughout the country in grades 1, 3, 6, 9, and 12. In the analysis that follows, grade 3 is omitted because the results of the achievement test of that grade have been thrown into a questionable light. The survey sample consisted of a 5 percent sample of schools and the information was comprehensive in that factual information was collected on the students' home background, such as education and occupation of parents.

The educational attainment of parents was listed in eight categories ranging from zero-to-seven-years-of-education to five-plus-years-of-college. It is possible to infer earnings from the Coleman data on parent's educational attainment and the 1960 U.S. Census of Population data on earnings by age and educational attainment of men and women.

The school outcome variable used in the analysis is verbal achievement,[9] and the learning model is statistically estimated for grades 1, 6, 9, and 12, white and nonwhite students separately.

TABLE 3

Linear Education Model, Student Verbal Achievement As A Function Of The Educational Services Of Parents And School Resources: Grade 1,6,9, And 12, White and Nonwhite Regressions

$$A_i = \alpha_{i0} + \alpha_{i1}M + \alpha_{i2}F + \alpha_{i3}S + U_i$$

Grade Level, White and Nonwhite	Regression Coefficients (t-statistics in parentheses) Mother's Educational Services	Father's Educational Services	School Resources	Constant	$\lvert X'X \rvert$	R^2
Grade 1 (White)	.0003 (153.70)	.0019 (147.90)		16.0606	.058	.888
Grade 1 (Nonwhite)	.0003 (100.24)	.0033 (153.14)		14.2593	.075	.875
Grade 6 (White)	.0005 (144.62)	.0023 (141.67)	.0335 (63.01)	-38.7431	.087	.792
Grade 6 (Nonwhite)	.0002 (30.95)	.0017 (37.68)	.0168 (74.55)	-10.8825	.246	.499
Grade 9 (White)	.0007 (362.66)	.0026 (354.88)	.0048 (21.77)	3.5223	.043	.914
Grade 9 (Nonwhite)	.0005 (138.96)	.0022 (116.57)	.0033 (42.05)	4.0680	.094	.785
Grade 12 (White)	.0006 (339.77)	.0018 (333.77)	-.0020 (-13.20)	33.4004	.055	.902
Grade 12 (Nonwhite)	.0005 (150.02)	.0018 (103.35)	.0015 (20.56)	9.1227	.095	.791

Notation:

R^2 - Coefficient of determination

$\lvert X'X \rvert$ - determinant * of zero - order correlation matrix

* This determinant is a test for the presence of multicollinearity. If the determinant approaches zero there is evidence of severe multicollinearity. When it is equal to unity, perfect orthogonality exists.

TABLE 4

The Marginal Rates of Substitution Among Parental Educational Services and School Resources for Linear Model: Grade 1, 6, 9, and 12, White and Nonwhite Students

Grade Level, White and Nonwhite		Marginal Rates of Substitution α_2/α_1	α_3/α_1	α_3/α_2
Grade 1	(White)	6.3		
Grade 1	(Nonwhite	11.0		
Grade 6	(White)	4.6	67.0	14.6
Grade 6	(Nonwhite)	8.5	84.0	9.9
Grade 9	(White)	3.7	6.9	1.8
Grade 9	(Nonwhite)	4.4	6.6	1.5
Grade 12	(White)	3.0	-3.3	-1.1
Grade 12	(Nonwhite)	3.6	3.0	.8

Notation:

α_2/α_1 - Marginal rate of substitution (linear model) between fathers' and mothers' contributed educational services

α_3/α_1 - Marginal rate of substitution between school resources and mothers' educational services

α_3/α_2 - Marginal rate of substitution between school resources and fathers' educational services

The Empirical Results

The results of the empirical estimation[10] of the linear model are presented in Table 3. The regression coefficients and coefficients of determination are presented there for four grades and for whites and nonwhites. The amount of variation explained by the educational services of parents and school resources is substantial, generally falling within the 80 to 90 percent range. The stability over grades and races of the parental coefficients should be noted, whereas the influence of school resources over the grades is less stable. The impact of the school upon achievement appears to be the greatest at grade 6, for both whites and nonwhites, and it diminishes at grade 9 and again at grade 12, where it actually has a negative influence upon the verbal achievement of white students. This negative coefficient is most likely a statistical artifact caused by the multicollinearity among the independent variables.

The linear relationship in equation (5) implies a constant marginal product of parental and school investments regardless of the intensity of these services. The absolute magnitude of the marginal product may not be a valid analytical concept in this context because the achievement measure is not a cardinal one, there being no well-defined unit of measurement of achievement. The marginal rate of substitution among alternative educational investments may provide us with useful information of parental and school services. The marginal rate of substitution is an indication of the efficiency of one set of educational resources relative to another.

The marginal rate of substitution among investments is the ratio of the marginal product of one investment to the marginal product of another. In our linear model, as specified in equation (5), the marginal rate of substitution between the school and the mother is defined as:

$$(8)\quad \frac{\delta A_i/\delta S_i}{\delta A_i/\delta M_i} = \frac{\alpha_{i3}}{\alpha_{i1}}$$

This ratio indicates the marginal contribution of school investments relative to the mother's contributed educational services. This ratio for each grade level and the ratios for the father to the mother and for the school to the father are presented in Table 4.

α_2/α_1, the marginal rate of substitution between the contributed educational services of the father and the mother, is greater than 3 at each grade level, for the white and nonwhite alike. The magnitude of the ratio, however, is much different between the grades. The ratio is high for the first and sixth grade nonwhite students. This indicates the substantial amount of influence the presence of the nonwhite male in the home seems to have.

The ratio α_3/α_1 shows a diminishing influence of school investments relative to the mother's contribution over the educational life of the student. The α_3/α_1 ratio shows the marginal contribution of schools to be much greater than that of the mother. The marginal rate of substitution between the school and the father is near unity for the ninth and twelfth grades.

If the linear model is a good predictor of achievement, then we do glean some insights into the relative efficiencies of the parental and school investments. The father and the school are relatively more productive than the mother in contributing educational services. On the other hand, the mother's contributed educational services may be over-extended; thus, her relatively small influence upon achievement. If the mother is over-extended, one must find a way of making her activity more effective in the years she provides educational services.

The linear formulation of the model does not address itself to the investment intensity problem. In fact, it is this aspect of the linear model that suggests a more refined formulation of the model in which intensities do play a role. Furthermore, a linear model, with its implicit assumption of infinite substitution among the investment alternatives, is based on very tenuous grounds. This has an unrealistic policy implication of choosing the investment alternative that has the largest marginal product and channelling all funds to that particular investment. Such a strategy is neither theoretically satisfying nor practical within the current home-school structure of our society.

The results of the logarithmic formulation of the achievement model are presented in Table 5. On all counts with the single exception of grade 6, nonwhite students, this model has a better statistical fit as measured by higher R^2's. The coefficient of determination is but one criteria for judging the goodness-of-fit of these alternative models. There is reason to believe that increases in investments will yield different changes in the achievement depending on their timing. It may be relatively easy to move forward on the lower range of the scale, even assuming that there are no measurement errors in the output variable. Whereas, at higher levels on that scale, movements induced by increased investments may be hard to come by. More importantly, the logarithmic formulation of the model specifies that investment intensities are influential in determining the marginal products of these alternative education investments of parents and school.

Again, because of the uncertain nature of the measurement characteristics of the dependent variable, the size of the marginal product of an investment must be interpreted with care; yet the ratio of the marginal products of two investments is still a valid analytical concept. From equation (7) it is possible to derive ratios of marginal products:

TABLE 5

Nonlinear Education Model, Student Verbal Achievement As A Function of the Educational Services of Parents and School Resources: Grade 1, 6, 9, and 12, White and Nonwhite Regressions

$$\ln A_i = \ln \beta_{10} + \beta_{11} \ln M + \beta_{12} \ln F + \beta_{13} \ln S + V_i$$

Grade Level, White and Nonwhite	Regression Coefficients (t-statistics in parentheses) Mother's Educational Services	Father's Educational Services	School Resources	Constant	$\lvert X'X \rvert$	R^2
Grade 1	.0761	.0781		1.7808	.026	.950
(White)	(256.39)	(205.67)				
Grade 1	.0652	.0801		1.8112	.055	.898
(Nonwhite)	(129.29)	(136.28)				
Grade 6	.1254	.1370	1.2929	-8.3796	.046	.878
(White)	(232.31)	(174.75)	(53.39)			
Grade 6	.0624	.0787	1.1256	-6.4015	.208	.460
(Nonwhite)	(37.42)	(29.26)	(53.47)			
Grade 9	.1989	.2289	.1238	-1.1169	.026	.942
(White)	(468.89)	(407.19)	(6.73)			
Grade 9	.1850	.1721	-.0914	.9012	.062	.805
(Nonwhite)	(180.27)	(112.57)	(-7.67)			
Grade 12	.1639	.1930	-.7056	6.4826	.043	.918
(White)	(387.84)	(355.78)	(-38.24)			
Grade 12	.2036	.1875	-.3634	2.9977	.065	.797
(Nonwhite)	(170.34)	(106.52)	(-25.23)			

Notation:

R^2 - Coefficient of determination
$\lvert X'X \rvert$ - determinant of zero - order correlation matrix

TABLE 6

The Marginal Rates of Substitution Among Parental Educational Services and School Resources For Logarithmic Model: Grades 1, 6, 9, and 12, White and Nonwhite Students

Grade Level, White and Nonwhite		Marginal Rates of Substitution $\beta_2 M/\beta_1 F$	$\beta_3 M/\beta_1 S$	$\beta_2 F/\beta_2 S$
Grade 1	(White)	5.7		
Grade 1	(Nonwhite)	6.7		
Grade 6	(White)	3.6	36.9	10.2
Grade 6	(Nonwhite)	4.5	40.4	9.1
Grade 9	(White)	3.2	1.6	.5
Grade 9	(Nonwhite)	2.8	-.8	-.3
Grade 12	(White)	2.97	-8.9	-3.0
Grade 12	(Nonwhite)	2.5	-2.3	-.9

Notation:

$\beta_2 M/\beta_1 F$ - Marginal rate of substitution (logarithmic model) between father's and mother's educational services

$\beta_3 M/\beta_1 S$ - Marginal rate of substitution between school resources and mother's educational services

$\beta_3 F/\beta_2 S$ - Marginal rate of substitution between school resources and father's educational services

$$\text{(9)} \quad \frac{\delta A_i / \delta S_i}{\delta A_i / \delta M_i} = \frac{\beta_{3i} M_i}{\beta_{1i} S_i}$$

and

$$\text{(10)} \quad \frac{\delta A_i / \delta F_i}{\delta A_i / \delta M_i} \quad \frac{\beta_{12} M_i}{\beta_{11} F_i}$$

Table 6 presents these ratios derived from the logarithmic model where the values of the independent variables are mean values.

The marginal rates of substitution in Table 6 are substantially lower than the marginal rates derived from the linear model, except for the cases where the school coefficient is negative. This is a strong indication of the model specification bias when a linear form is imposed upon a non-linear system. This result is another piece of evidence favoring the non-linear model because it is an indication that factor intensities are prevalent. The better fit plus the uniformly smaller marginal rates of substitutions represent strong evidence in favor of the non-linear model.

The marginal rates of substitution for the non-linear model have the same characteristics as the rates derived from the linear model. The father's contributed educational services are relatively more efficient than the mother's, although this relative efficiency diminishes over the educational life of the student. School resources appear to be more efficient than the mother's contributed services. For grade 6, the school is relatively more efficient than the services of fathers, and this is reversed for the white students at grade 9.

CONCLUSIONS AND POLICY IMPLICATIONS

This paper presented a new model of scholastic achievement that postulated a theory of learning in which the output of schools, student achievement, depended upon past as well as current educational investments which include the contributed educational services of parents and formal education investments. The model was empirically estimated and the results were highly significant in terms of high $\underline{R}^2$'s and regression coefficients. The achievement model was tested statistically in two mathematical forms, and the non-linear model produced the "best" fit. The empirical results show that the parental investments, as determined by the opportunity costs of the parents' time, are highly significant determinants of scholastic achievement.

This new approach of the determination of scholastic achievement presents an analytical tool to investigate the trade-off between school and parental investments. Such a trade-off is the foundation of compensatory education, for it gives an indication of the amount of school resources required to elevate a culturally deprived child to an achievement level comparable to the national norm.

It is worth noting that in grade 9, if the mean values of the contributed educational services of parents and school resources of white students are put in the nonwhite model, 55.8 percent of the gap between white and nonwhite verbal achievement is eliminated. This is another indication that the potential contribution to educational achievement of the improvement in the economic status of a student's family should not be overlooked by policy makers.

This paper points out that the concept of "equality of educational opportunity" that is currently in vogue is in need of close scrutiny. If it means the quality of current educational expenditures per student, the goal may be achieved simply by designing allocation formulas to ensure such equality. If, however, "equality of educational opportunity" refers to equal access to scholastic achievement, the costs of the compensatory education necessary to facilitate such equality are substantial. The "Great American Dream" does not come cheaply.

REFERENCES

1. Bartell, Ernest J., Costs and Benefits of Catholic Elementary and Secondary Schools. (Notre Dame: University of Notre Dame Press, 1969).

2. Becker, Gary, "A Theory of the Allocation of Time." The Economic Journal, Volume LXXV, September, 1965, pp. 493-517.

3. Becker, Gary, Human Capital: A Theoretical and Empirical Analysis with Special Reference to Education. (New York: Columbia University Press, 1964).

4. Bowles, Samuel, "Educational Production function." Final Report on Research Supported by U.S. Office of Education, February, 1969, Mimeographed

5. Bowles, Samuel S., and Levin, Henry M., "More on Multicollinearity and the Effectiveness of Schools." The Journal of Human Resources, Volume 3, No. 3 (Summer 1968), pp. 393-400.

6. Coleman, James S., et al. Equality of Educational Opportunity. Two Volumes. (Washington, D.C., Government Printing Office, 1966).

7. Hess, Robert D., and Shipman, Virginia D., "Maternal Influences Upon Early Learning," in Early Education. (Chicago: Aldine Publishing Company, 1968).

8. Jones, J.B., Lundsteen, Sara W., and Michael, W.B., "The Relationship of the Professional Employment Status of Mothers to Reading Achievement of Sixth-Grade Children," California Journal of Educational Research, 1967, 18, pp. 102-108.

9. Ribich, Thomas, Education and Poverty. (Washington, D.C., The Brookings Institution, 1968).

APPENDIX A

Allocation of Mother's Household Time

The allocation of the time used by the housewife is divided among six basic homemaking activities: (1) meal preparation, (2) dishwashing, (3) house care, (4) washing, (5) ironing, and (6) the physical care of the children and other family members. A General Electric Survey[11] indicates that a nonworking wife with children under six years old spent 1,135.16 hours per year in the physical care of children. These 21.8 hours per week compares to three hours per week spent on physical care of children by the housewife of a family with children between six and eighteen years. Estimates of the weekly time spent in homemaking by full-time homemakers vary between 50 and 52 hours per week.[12] Therefore, in a 50-hour work week of mothers, 43 percent of her time is devoted to the physical care of children.

This estimate of the percent of time devoted by the housewife to the physical care of children may, in fact, understate the value of the educational services provided by the mother. Manual tasks such as meal preparation may, in fact, present situations in which joint production may take place. While preparing meals, it is possible to maintain communication with children and also to present stimulus-response situations by demonstrating techniques to the pre-schooler. Examples of such joint production are too numerous to mention, and they do indicate that our estimate of the percent of the household work time of the mother may, in fact, be understated. We will not enter upon the delicate operation of separating such joint production and we will, therefore, content ourselves with the conservative estimate derived above.

APPENDIX B

Children in the Family and Child Spacing

The a's, b's, and c's are parameters appropriately defined as "the effective proportion of time devoted to the individual child." For the first three years of the child's preschool period, a_t is assumed to be the entire 43 percent of the mother's time. This assumption is based upon the vital statistics concerning the spacing of children where the average spacing between the first and second child is two and one-half years. The median intervals between births are given in Table B-1.

TABLE B-1

Median Interval Between Births by Color: 1959*

	Interval (1955-59)
White	
Median interval from--	
First marriage of mother to birth of first child	16.2
Birth of first child to birth of second child	28.2
Nonwhite	
Median interval from--	
First marriage of mother to birth of first child	11.9
Birth of first child to birth of second child	23.4

*Statistical Abstract of the United States: 1966, p. 52.

For the remaining preschool years, it is assumed that the mother devotes 21½ percent of her time to the child.

b_t coefficient reflects the combination of two factors that limit the amount of time the mother devotes to a particular child. First, once the child is in school, the amount of physical care by the mother is reduced to three hours per week for all the children (see Appendix A); and second, the mother divides her three hours among her children. The average number of children by the educational attainment of the mother is given in Table B-2. b_t is therefore determined by the two factors mentioned above for mothers with various levels of educational attainment.

TABLE B-2

Number of Children Ever Born Per 1,000 Women Ever Married, by Years of School Completed, 1960**

Years of school completed:		
Elementary:	Less than 8 years	3,091
	8 years	2,637
High School:	1 to 3 years	2,470
	4 years	2,074
College:	1 to 3 years	1,965
	4 years or more	1,704

**Statistical Abstract of the United States: 1966, p. 51.

In calculating the "opportunity cost" of the father's educational services, it is assumed that the father devotes one-twentieth of his time to the educational activities of his children and he spreads this time among his children. Therefore, c_t is determined by this time element and the number of children according to the level of his educational attainment.

APPENDIX C

Empirical Estimation Technique and Derivation of School Resource Index

The basic data used for testing the achievement model are derived from two sources, the Equality of Educational Opportunity Survey and 1960 Census of Population.[13] The EEOS data provide us with verbal achievement scores and the educational attainment of mother and father. The eight years-of-school-completed categories for parents ranged from 0-7 years to 5+ years of college, and these were translated into income figures with the 1960 Census data relating educational attainment to income. The eight categories of educational attainment for each parent provides us with 64 cells in which all students were categorized; and the mean achievement and frequency was computed for each cell. This was the empirical foundation of the weighted multiple regressions that were performed for grades 1, 6, 9, and 12, white and nonwhite separate.

Accumulated school expenditures per student are weighed 1960 current expenditures per student[14] derived from State data on expenditures and weighted across States by the number of men or women of a given level of educational attainment. Although school expenditures do vary drastically within any given State, enough differences do exist among States so that the variation is adequate to give the variable meaning and to portray differences that exist in expenditures per pupil.

TABLE C-1

Means, Standard Deviations, and Number of Observations of Variables[a] in Linear Model: By Grade Level and Race

	Mean	Standard Deviation	Number of Observations
Grade 1 (White)			18,861
Verbal Achievement	19.17	1.07	
Mother	4991.46	2032.44	
Father	933.22	286.79	
Grade 1 (Nonwhite)			12,453
Verbal Achievement	16.41	.96	
Mother	2507.78	1471.45	
Father	443.99	178.80	
Grade 6 (White)			40,738
Verbal Achievement	36.80	3.72	
Mother	7751.83	3070.00	
Father	2268.03	736.65	
School	1982.90	19.55	
Grade 6 (Nonwhite)			21,034
Verbal Achievement	23.71	3.31	
Mother	4901.34	3070.21	
Father	1283.87	503.59	
School	1871.14	86.95	
Grade 9 (White)			65,376
Verbal Achievement	33.73	4.91	
Mother	8957.17	3407.64	
Father	3136.80	1045.66	
School	3228.08	29.49	
Grade 9 (Nonwhite)			28,958
Verbal Achievement	20.71	3.16	
Mother	5561.27	3092.21	
Father	1800.53	660.62	
School	3059.01	137.07	
Grade 12 (White)			57,944
Verbal Achievement	37.30	4.54	
Mother	10727.66	4118.90	
Father	4181.73	1438.98	
School	4772.35	42.67	
Grade 12 (Nonwhite)			22,821
Verbal Achievement	23.60	3.77	
Mother	6626.70	3718.65	
Father	2346.75	906.67	
School	4506.88	202.70	

[a]Variables include verbal achievement of student, parental contributed educational services, and school resources.

*The author is a Public Policy Fellow at the Brookings Institution and Assistant Professor of Economics, The University of Notre Dame.

The author wishes to thank H. Levin and H. Grabowski for comments on an earlier version of this paper. Special thanks goes to Mrs. Judith Zinter of the Office of Program Planning and Evaluation, U.S.Office of Education, for performing the programming and computations and for comments at various stages in the development of this paper.

1.See Hess and Shipman(7) and Jones, Lundsteen, and Michael(8), for example.

2.The opportunity cost of the parent in the labor market is not the ideal criteria for valuing parental educational services. The true measure of the value of educational services would be the price of specific parental services that influence achievement of the child. The current state of educational research, however, does not provide enough evidence to construct a set of implicit prices for the educational services of parents. Therefore, the "opportunity cost" concept, with its implicit assumptions of perfect labor market mobility of parents and perfect flows of market information, will be used to approximate the true prices of the educational services of parents.

3.See Appendix A for the complete breakdown of the activities and the number of hours per week devoted to the physical care of children.

4.See Jones, Lundsteen and Michael(8).

5.See Becker(2).

6.See Appendix B.

7.Per pupil expenditures of schools was also accumulated in the same manner as parental investments.

8.All dollar figures in constant 1966 dollars.

9.The criticisms of the shortcomings of the Coleman survey are too numerous to mention. Errors in the measurement of variables are present in the data, and results should be interpreted with care.

10.See Appendix C on data and the empirical estimation technique.

11.General Electric Company, Amount of Time Spent in House Work, Small Appliance Division Survey, Bridgeport, November 24, 1952, pp. 20-30.

12.Jean Warren, "Time Resource or Utility?", Journal of Home Economics, Vol. 49, No. 1, (January, 1957), pp. 20-22.

13.U.S. Bureau of the Census, Census of population: 1960, (Washington: U.S. Government Printing Office, 1963), PC(2), 5B, "Educational Attainment," Tables 6 and 7.

14.Digest of Educational Statistics: 1963, (Washington: U.S. Government Printing Office, 1963), page 46.

VIII

FIELD EXPERIMENTATION IN INCOME MAINTENANCE

Chairman, HAROLD WATTS, University of Wisconsin

A MODEL FOR OPTIMIZING EXPERIMENTAL DESIGNS FOR ESTIMATING RESPONSE SURFACES

John Conlisk, University of California, San Diego
Harold Watts, University of Wisconsin

I. INTRODUCTION

Among many problems and challenges coming from the Graduated Work Incentive Experiment, one in particular has provided an opportunity for the authors, as econometricians, to round out their statistical education; and, perhaps, to make a contribution to the literature on experimental design. The problem we faced is a familiar economic one--how to get the most of some desirable output from limited inputs of financial and other resources, while observing various additional constraints.

For the graduated Work Incentive Experiment, the desirable output was precision in estimating the effects of alternative income maintenance policies. The principal scarce input was money; but, because of earlier choices regarding the scale and structure of the experiment, the number of families included in the experiment was also treated as a scarce input. Among the additional constraints on the problem were bounds on the guarantee levels, marginal tax rates, and pre-experiment income levels of the families in the experiment. It was also decided that an index of policy-makers' interest in alternative income maintenance policies should affect the design. These factors, together with specifications of an appropriate behavioral response function, have been fitted into a tractable mathematical model from which optimal allocation of the scarce experimental funds and families can be derived.

The model itself is not specific to the Graduated Work Incentive Experiment; rather it may be viewed as a general experimental design model for regression analysis. The model is both operational and quite flexible. It is operational in the sense that it reduces to a problem of minimizing a convex non-linear objective function subject to a set of linear constraints, a problem for which efficient computer solutions are available. The model is flexible in the sense that it can handle problems of arbitrary size, arbitrary regions of observation, alternative response functional forms, multiple response functions, varying experimental objectives, unequal costs per observation, unequal error variances, multiple constraints on the design, and other variations.

Section II sets out the mathematical bare bones of the model. Section III discusses the application of its various components to the Graduated Work Incentive Experiment.

II. THE MATHEMATICAL MODEL

Suppose an N-observation sample must be designed for estimting a response function

$$(1)\quad y_r = f(z_{r1},\ldots,z_{rk}) + e_r \quad r = 1,\ldots,N$$

where the context is as follows. The z_{ij} are design variables; they are subject to experimental control either by stratification (as when families of given income level are chosen in a tax experiment) or by direct control (as when payment levels are set in such an experiment). The y_r are the observable responses; and the e_r are (at least partially) unobservable random errors. The cost of a given observation may vary with the levels of $(z_{r1},\ldots,z_{rk})$ for that observation; and there is a maximum budget C that can be spent. Observations are restricted to a given region in the k-dimensional design space of the design variables. Finally, the function f is linear in parameters, and tractable assumptions for the e_r are allowed; so (1) is a standard regression equation.

A. The Basic Model

The building blocks of the basic model are 1) the regression model, 2) the admissable regressor rows, 3) the objective function, and 4) the budget constraint.

1. The regression model for equation (1) is

$$(2)\quad \begin{aligned} y &= X\beta + e \\ E(e) &= 0 \qquad V(e) = \sigma^2 I \\ b &= (X'X)^{-1}X'y \qquad V(b) = \sigma^2(X'X)^{-1} \end{aligned}$$

where y is the dependent variable vector, X the regressor variable matrix, e the error vector, β the coefficient vector, b the least squares estimate of β, and V(.) the variance matrix operator. It is assumed that X is of full column rank. The regressor matrix X depends, row for row, on the design matrix $Z = [z_{ij}]$, but is not in general the same. For instance, with k = 2, it might be that $(x_{r1}, x_{r2}, x_{r3}) = (1, z_{r1}, \log(z_{r1}/z_{r2}))$.

2. The admissable regressor rows may be introduced as follows. The design problem is to choose Z, and thus X, in some optimal way. Each row of Z represents an observation on the design variables. Complete freedom in choosing rows is not allowed. Instead, it will be assumed that there are a fixed number of admissable rows, each of which may be represented a number of times in Z. Corresponding to each admissable row for Z is an admissable row for X. Let m be the number of such admissable rows for X, x_i be the ith of them, and n_i be the number of times x_i is represented in X. Then X is composed of n_1 rows like x_1, n_2 rows like x_2, and so on. The total sample size N and the regressor cross product matrix X'X are given by

$$(3)\quad N = \sum_{i=1}^{m} n_i \qquad X'X = \sum_{i=1}^{m} n_i x_i' x_i \,.$$

So the design problem of choosing Z, and thus X, is simplified to the problem of choosing the m non-negative integers $n_1,\ldots,n_m$, given a set of admissable rows. The admissable rows for the design matrix Z represent points in the design space, called design points. The experimenter chooses these design points so as to give the relevant region of the design space adequate coverage. For X to be of full rank, the x_i, when stacked into an m-row matrix, must be of full rank (and, as discussed below, an appropriate number of the n_i must be positive).

3. An objective function to optimize in choosing the n_i is required. Suppose the experimenter's goal is accurate estimation of a vector $P\beta$ of linear combinations of the elements of β. The best linear unbiased estimate of $P\beta$ is Pb. It is assumed that the experimenter wishes to minimize a weighted sum of the variances of the elements of Pb. This objective function may be written tr[WV(Pb)], where tr(.) is the trace operator and W is a diagonal weight matrix whose diagonal elements indicate the policy importances to the experimenter of the elements of $P\beta$. Substituting from (2) and (3) and multiplying by the constant σ^{-2} gives the objective function as used, call it ϕ. Letting D = P'WP,

$$\begin{aligned}\phi(n_1,\ldots,n_m) &= \sigma^{-2}\mathrm{tr}[WV(Pb)]\\ &= \sigma^{-2}\mathrm{tr}[P'WPV(b)]\\ &= \sigma^{-2}\mathrm{tr}[P'WP\sigma^2(X'X)^{-1}]\\ &= \mathrm{tr}[D(\sum_{i=1}^{m} n_i x_i' x_i)^{-1}]\ .\end{aligned}$$

4. The budget constraint of the basic design model is $\Sigma_{i=1}^{m} c_i n_i \leq C$, where c_i is the cost of one observation at the ith design point (that is, with regressor row x_i) and C is the total available budget.

Given these building blocks, the basic design model may be simply stated as follows. With D = P'WP,

$$\text{minimize} \quad \phi(n_1,\ldots,n_m) = \mathrm{tr}[D(\sum_{i=1}^{m} n_i x_i' x_i)^{-1}] \tag{4}$$
$$\text{subject to} \quad \sum_{i=1}^{m} c_i n_i \leq C, \quad n_1 \geq 0,\ \ldots,\ n_m \geq 0.$$

Strictly speaking, this is an integer programming problem, since the n_i are integers. Practically speaking, however, little will be lost in practice by treating the n_i as continuous in solving (4) and then rounding off. Economists may see the design problem (4) as analagous to a utility-maximization problem from consumer choice theory, where ϕ corresponds to an inverse measure of utility, the n_i to amounts of m goods, the c_i to prices, and C to available income.

Summarizing then, the basic design model for regression analysis requires the experimenter to specify 1) a regression model, 2) a set of design points and the corresponding regressor rows x_i, 3) two objective function matrices P and W, and 4) the costs c_i and budget C. Then he must solve the programming problem (4).

B. Further Discussion of the Model

1. Consider the derivatives and convexity of $\phi = \phi(n_1,\ldots,n_m)$. Let $S = (\Sigma_{i=1}^{m} n_i x_i' x_i)^{-1}$. Then

$$\begin{aligned}\partial\phi/\partial n_i &= \partial\,\mathrm{tr}(DS)/\partial n_i = \mathrm{tr}(D\partial S/\partial n_i)\\ &= -\,\mathrm{tr}(DSx_i'x_iS) = -\,\mathrm{tr}(x_iSDSx_i')\\ &= -\,x_iSDSx_i'\ ,\end{aligned} \tag{5}$$
$$\begin{aligned}\partial^2\phi/\partial n_i\partial n_j &= -x_i(\partial S/\partial n_j)DSx_i' - x_iSD(\partial S/\partial n_j)x_i'\\ &= (x_iSx_j')(x_jSDSx_i') + (x_iSDSx_j')(x_jSx_i')\\ &= 2(x_iSx_j')(x_iSDSx_j')\end{aligned}$$

Since $SDS = (W^{1/2}PS)'(W^{1/2}PS)$ is non-negative definite, then the first partials $\partial\phi/\partial n_i$ are non-positive, as expected; an increase in the sample can do no harm. (Note that $\partial\phi/\partial n_i$ can be zero, since D may be of less than full rank.) The convexity of ϕ can be proved by showing the matrix of second partials to be non-negative definite. Letting $\otimes$ denote Kronecker multiplication, $\partial^2\phi/\partial n_i\partial n_j$ can be conveniently restated

$$\begin{aligned}\partial^2\phi/\partial n_i\partial n_j &= 2(x_iSx_j')\otimes(x_iSDSx_j')\\ &= 2(x_i\otimes x_i)[S\otimes(SDS)](x_j\otimes x_j)'.\end{aligned}$$

Hence the entire matrix of second partials may be written

$$2\begin{pmatrix} x_1\otimes x_1 & & O\\ & \ddots & \\ O & & x_m\otimes x_m\end{pmatrix} \times \begin{pmatrix} S\otimes(SDS) & \cdots & S\otimes(SDS)\\ \vdots & & \vdots\\ S\otimes(SDS) & \cdots & S\otimes(SDS)\end{pmatrix} \times \begin{pmatrix} x_1\otimes x_1 & & O\\ & \ddots & \\ O & & x_m\otimes x_m\end{pmatrix}' \tag{6}$$

This product is non-negative definite if the central matrix is. But the central matrix may be alternately written $(uu')\otimes S\otimes(SDS)$ where u is a column of ones; and $(uu')\otimes S\otimes(SDS)$ is non-negative definite because the Kronecker product of non-negative definite and positive definite matrices is non-negative definite.

2. Solving the design problem (4) is not difficult since the Kuhn-Tuchker first order minimization conditions take the simple form

$$\begin{aligned}(\partial\phi/\partial n_i)/c_i &= \lambda \text{ for all i with optimal } n_i > 0,\\ (\partial\phi/\partial n_i)/c_i &> \lambda \text{ for all i with optimal } n_i = 0.\end{aligned} \tag{7}$$

Equations (7) say that all design points included positively in the optimal design have the same marginal effectiveness per dollar of cost, call it λ, in reducing ϕ; while all design points excluded have lesser effectiveness. λ is the shadow price of the budget constraint; it is negative and equals $\partial\phi/\partial C$, evaluated at the optimum. The budget constraint will of course hold with equality. The first order conditions (7) indeed assure a global minimum since ϕ is convex; though the minimum may not be unique, since ϕ is not strictly convex. A simple iterative solution procedure may be based on the idea of letting the relative sizes of the $(\partial\phi/\partial n_i)/c_i$ determine how the n_i shift up and down from iteration to iteration.

3. Four useful <u>scale properties</u> of the model may be stated. Noting first that the objective function ϕ is homogeneous of degree minus one in the n_i, consider the relations

$$\begin{aligned} n_i &= \alpha_i N , \\ (8) \quad N &= C/\sum_{i=1}^{m}\alpha_i c_i , \\ \phi &= (1/C)(\sum_{i=1}^{m}\alpha_i c_i)\mathrm{tr}[D(\sum_{i=1}^{m}\alpha_i x_i' x_i)^{-1}]. \end{aligned}$$

The first of these defines the fractions α_i, which give the proportional allocation of the total sample N over the design points. The second uses the first to rewrite the budget constraint (with equality holding). The third uses the first two to rewrite the objective function. The scale properties are: <u>First</u>, for a given proportional allocation $(\alpha_1,\ldots,\alpha_m)$, a change in C will cause an equiproportionate change in ϕ. <u>Second</u>, for given α_i, an equiproportionate change in all the c_i will result in the same proportionate change in ϕ, and the same inversely proportionate changes in N and the n_i. <u>Third</u>, for given α_i, equiproportionate changes in C and all the c_i will leave ϕ, N, and the n_i unchanged. <u>Fourth</u>, the optimal α_i are independent of the value of C.

4. In <u>constructing the P-matrix</u>, there are many sensible choices the experimenter might make. Two rather neutral examples are

$$(9) \qquad P = I \quad \text{and} \quad P = \begin{pmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_m \end{pmatrix} .$$

The first choice would imply that the experimenter was interested in estimating the elements of $P\beta = \beta$ themselves. The second choice would imply that he was interested in estimating the heights of the response function over the m design points. In making the choice, no firm constraint need be put on the number of rows in P, though several things about this may be noted. If P is not of full row rank, then it may always be condensed row-wise until it is of full row rank; so P need never have more rows than columns. That is, a P_o with full row rank and a corresponding W_o may always be found such that $P'WP = P_o'W_oP_o = D$ in (4), regardless of the rank or the number of rows in P. If the condensed matrix P_o has fewer rows than columns, there can be trouble. For example, suppose the one-row P-matrix $P = P_o = x_1$. Choice of this P would imply that the experimenter was solely interested in estimating the height $P\beta = x_1\beta$ of the response function over the first design point. In this case, the optimal design would be to put all observations at the first design point; so $n_1 = N$ and $n_2 = \ldots = n_m = 0$. This would mean, for an x_1 with more than one element, that the inverse in (4) would not exist; so the objective function would break down. This breakdown is not necessary whenever P_o has fewer rows than columns; it may or may not happen, depending on the values of P_o and the x_i. When it does happen, it is essentially because the experimenter has specified a regression form more complicated than he really wishes to estimate, as indicated by his choice of P. So the solution is to simplify the regression form and/or to increase the row rank of P. For either of P-specifications (9), of course, the compacted matrix P_o is non-singular; so the breakdown will not occur. (The P_o of this paragraph is only for discussion purposes; it need not be solved for in practice.)

5. An <u>explicit solution for a one-way analysis of variance model</u> is available. Suppose the x_i take the form

$$x_1 = (1,0,\ldots,0), \quad x_2 = (0,1,\ldots,0), \quad \ldots, \quad x_m = (0,0,\ldots,1).$$

Then the model is a one-way analysis of variance model where β_i and n_i are the mean of and the number of observations allocated to the ith cell, respectively. Letting P = I as in the first of examples (9), and letting w_i be the ith diagonal element of W, the explicit solution to (4) for the optimal n_i is

$$(10) \quad n_i = (w_i/c_i)^{1/2}C/\Sigma_j(w_jc_j)^{1/2}$$
$$i = 1,\ldots,m.$$

6. <u>Orthogonality of the regressor matrix X</u> is not an optimality condition for the design model of this paper; though there is a well known theorem in the design literature giving conditions under which orthogonality is optimal. It is useful to review this theorem and see why the conditions may not be met. For this discussion, suppose the columns of X are not functionally related; so orthogonality of X is at least possible. (For example, one column of X may not contain square terms of another column.) Further suppose the elements of X are expressed as deviations from a point in the center of the region of interest (except possibly for a column of ones). Now consider three assumptions. <u>First</u>, all observations cost the same amount c; so the total sample is fixed at N = C/c. <u>Second</u>, the size of the region of interest is determined by putting upper bounds a_j on the mean squares of the regressor variables. (That is, for all j, suppose the jth diagonal element of X'X/N may not exceed a_j.)

Third, minimization of the variances of the elements of b is the objective. Given these assumptions, the following derivation due to Tocher (1952) applies. Letting T be an upper triangular matrix such that X'X = T'T, and letting a double subscript on a matrix symbol denote the corresponding element, then

$$
\begin{aligned}
\text{var}(b_j) &= \sigma^2(X'X)^{-1}_{jj} = \sigma^2(T'T)^{-1}_{jj} \\
&= \sigma^2(T^{-1}T'^{-1})_{jj} = \sigma^2\Sigma_i(T^{-1})^2_{ij} \\
&\geq \sigma^2(T^{-1})^2_{jj} = \sigma^2/T^2_{jj} \geq \sigma^2/\Sigma_i T^2_{ij} \\
&= \sigma^2/(T'T)_{jj} = \sigma^2/(X'X)_{jj} \geq \sigma^2/Na_j \, .
\end{aligned} \tag{11}
$$

It may be seen that the equalities in (11) will hold--and thus $\text{var}(b_j)$ will achieve its lower bound--if and only if X'X, T'T, and T are diagonal, and the $(X'X/N)_{jj}$ are pushed to their bounds a_j. So optimality in terms of minimum $\text{var}(b_j)$ requires orthogonality of X, given the three assumptions listed.

By seeing how the assumptions may not be met in the context of this paper, we may see how orthogonality may be non-optimal. First, the result will not hold in general if the cost of an observation varies with the values of the regressor variables. For example, if there are just two regressor variables (in addition possibly to a column of ones in X) as shown on Figure 1, and if the costs per observation are higher in the second and fourth quadrants than in the first and third quadrants, then the model will tend to set up a positive collinearity between the variables at the optimum. (A similar effect may result when further constraints (such as those discussed below) are added to the model.) Second, the result will not hold in general if the region of interest is determined in another way. The assumption above, by limiting each regressor variable separately, sets up a rectangular region like R_o in Figure 1; whereas the experimenter may in fact be faced by a region of interest

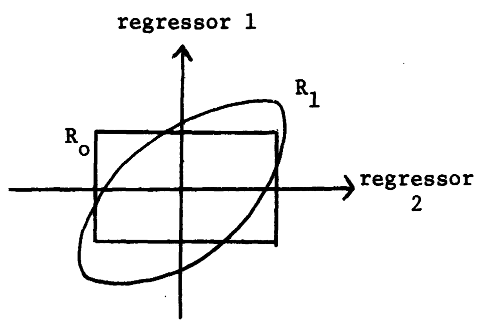


Figure 1

like R_1. In the latter case, the model will tend to set up a negative collinearity between the two regressor variables. Third, the result will not hold in general if the experimenter's objective is not minimization of the $\text{var}(b_j)$ separately. For example, the experimenter may be more interested in $\text{var}(b_1-b_2)$ than $\text{var}(b_1)$ and $\text{var}(b_2)$ separately; this interest can easily be expressed by appropriate choice of P. In such a case, the model would tend to set a negative correlation between regressor variables 1 and 2.

C. Extensions of the Model

The following extensions are considered singly, but they can easily be used in combination.

1. The choice of a specific regression functional form for the response function (1) is crucial to the character of the optimal design. For instance, if (1) were first degree polynomial, the optimal design would tend to concentrate only on design points around the boundary of the relevant design space region. If, however, (1) were third degree polynomial, the optimal design would tend to pick up some design points in the interior for identifying curvature. Unfortunately, the experimenter seldom knows ahead of time what the appropriate functional form is. Suppose the experimenter is considering q candidate functional forms, and suppose $\phi_j(n_1,\ldots,n_m)$ for $j = 1,\ldots,q$ are the corresponding objective functions. It is helpful to think of $\phi_j(n_1,\ldots,n_m)$ as a loss function where j indexes states of nature (functional forms) and $(n_1,\ldots,n_m)$ represents actions (sample designs). One way to investigate the sensitivity of the model to alternate functional forms is then to construct a qXq loss table, where the q states of nature are the q functional forms, and the q actions are the optimal designs for the q functional forms taken one at a time. In generating a design for ultimate use, the experimenter might use the Bayes strategy of minimizing the expected loss

$$
G(n_1,\ldots,n_m) = \sum_{j=1}^{q} \pi_j\phi_j(n_1,\ldots,n_m) \tag{12}
$$

where the π_j are prior probabilities for the states of nature, or functional forms. Replacing $\phi(n_1,\ldots,n_m)$ by $G(n_1,\ldots,n_m)$ in the design problem (4) above does not greatly complicate things. The sum of convex functions is convex; so G is convex. The derivatives (5) must be replaced by a π_j-weighted sums of expressions like (5). With appropriate minor modifications, all the other comments of the last section carry over.

2. Non-linearity in the parameters of the response function (1) may be simply handled by replacing the function with its Taylor series linearization about some guessed parameter values. This approach has been used by Box and Lucas (1959). Similarly, one might view the vector $P\beta$ as a linear approximation to some non-linear vector-valued function of β.

3. Multiple objectives in an experiment may be handled by a simple extension of the model. Suppose the experimenter wishes to estimate a set of response functions. (For instance, a medical experimenter might wish to estimate various physiological responses to controlled dieting.) Let the corresponding regression models be stated compactly as $Y = X\mathbf{B} + E$; where Y, $\mathbf{B}$, and E are matrices whose columns correspond to different regressions; and where the rows of E are independent with zero means and identical variance matrices σ^2U. It is known (see Goldberger (1964)

pp. 201-12, 246-8) that $B = (X'X)^{-1}X'Y$ is the best linear unbiased estimate of $\mathbf{B}$; and that $V(b_B) = \sigma^2 U \otimes (X'X)^{-1}$, where b_B is the vector gotten by stacking the columns of B in order from the first down to the last. By analogy to the objective function construction above, suppose the experimenter's estimate of interest is the vector Pb_B of linear combinations of the elements of b_B; and suppose he wishes to minimize the weighted sum $tr[WV(Pb_B)]$ of the variances of elements of Pb_B, where W is a diagonal weight matrix. The objective function, call if J, may then be written (with D = P'WP)

$$(13) \quad J(n_1,\ldots,n_m) = tr[D(U \otimes (\sum_{i=1}^{m} n_i x_i' x_i)^{-1})] \quad .$$

Replacing $\phi(n_1,\ldots,n_m)$ by $J(n_1,\ldots,n_m)$ in the design problem (4) does not substantively change the programming problem to be solved. The experimenter, in setting up the problem now has the one additional task of specifying a value for U. If U is diagonal (which seems unlikely in practice), then the various response functions to be estimated are independent; and (13) can be put in the form of (12) with the ϕ_j representing objective functions for the separate response functions, and the π_j equalling the diagonal elements of U.

4. Alternative forms for the objective function are available, typically based on some function of the variance matrix $V(b)=\sigma^2(X'X)^{-1}$ (that is, typically based on some function of the "information matrix" X'X). Two convenient possibilities are the determinant, or "generalized variance," function $|V(b)|$ and the trace function $tr(V(b))$ used here (neglect the weight matrix D momentarily). These functions are convenient because they are continuous, differentiable, convex, and so on. (See Kiefer (1959) for a discussion of these and other possibilities.) Minimizing $|V(b)|$ may be conveniently rationalized by noting that, if b is normal, the volume of a confidence ellipsoid around b, for given probability of containing β, is proportional to $|V(b)|$. Minimizing $tr(V(b))$ may be conveniently rationalized by noting that $tr(V(b)) = E(b-\beta)'(b-\beta)$ (the expected value of a quadratic loss function) or by noting, as above, that $tr(V(b))$ is the sum of the variances of the regression coefficient estimates. In a sense, $|V(b)|$ and $tr(V(b))$ are not very different objectives. Let $\lambda_1,\ldots,\lambda_s$ be the eigenvalues of V(b); they must all be positive. Then it is known that $|V(b)| = \lambda_1\cdots\lambda_s$ and $tr(V(b)) = \lambda_1+\ldots+\lambda_s$. So minimizing $|V(b)|$ is equivalent to minimizing the geometric mean of a set of positive numbers, and minimizing $tr(V(b))$ is equivalent to minimizing the arithmetic mean of the same numbers.

A determinant objective function is quite capable of carrying the design problems discussed here. However, the trace function is actually used because it is easier to weight. The trace function is a linear function of elements of V(b) and thus takes simple linear weights; whereas the determinant function is multiplicative and thus does not take linear weights. More specifically, it makes sense to replace $tr(V(b))$ by $tr(DV(b)) = E(b-\beta)'D(b-\beta)$; whereas it does not make sense to replace $|V(b)|$ by $|DV(b)|$. D must be square or $|DV(b)|$ is not defined; and D must be nonsingular or $|DV(b)| = 0$. But, with D nonsingular, $|DV(b)| = |D||V(b)|$; so minimizing $|DV(b)|$ is the same as simply minimizing $|V(b)|$. On the other hand, if the experimenter does not wish to use weights, and if computational ease is important, then $|V(b)| = |\sigma^2(X'X)^{-1}| = 1/|\sigma^{-2}X'X|$ may be more appropriate because it does not involve matrix inversion.

5. Unequal error variances in the regression model can easily be handled. Suppose the error variance differs from design point to design point such that the error variance corresponding to the ith regressor row x_i is $\sigma^2 v_i$. It may be shown that the objective function then becomes $\phi(n_1,\ldots,n_m) = tr[D(\Sigma_{i=1}^{m} n_i v_i^{-1} x_i' x_i)^{-1}]$, which introduces a very minor change indeed in the design problem (4). Of course, the experimenter must specify the v_i.

6. Attrition of observations from the sample can arise when, for example, some families which initially agree to be part of a cross-family sample later drop out; or when, for example, some observations in a laboratory experiment are unusable due to experimenter error. Suppose the attrition fraction varies by design point such that the attrition fraction corresponding to the ith regressor row x_i is μ_i. This can be easily handled in the design model (4) by replacing the n_i by $\mu_i n_i$ in the objective function. If attrition also affects the costs c_i, then they must be adjusted. For example, a family which drops out of a cross-family experiment may cost less than a family which stays the duration. If c_i' and c_i'' are the costs of observations which do and do not stay in the sample, respectively, then the appropriate cost to enter in the design model (4) is $c_i = (1-\mu_i)c_i' + \mu_i c_i''$.

7. Budget minimization subject to a maximum error constraint may sometimes be the experimenter's problem rather than error minimization subject to a maximum budget constraint. In such a case, the problem corresponding to (4) would be

$$\begin{aligned} &\text{minimize} \quad C = c_1 n_1 + \ldots + c_m n_m \\ (14) \quad &\text{subject to} \\ &\phi(n_1,\ldots,n_m) = tr[D(\sum_{i=1}^{m} n_i x_i' x_i)^{-1}] \le \phi_o \\ &n_1 \ge 0 \ , \ \ldots \ , \ n_m \ge 0 \end{aligned}$$

where ϕ_o is a pre-selected maximum admissable error. Given a solution procedure for the original problem (4), solution of the new problem (14) is easy. Note that the optimization conditions (7) apply to (14) as well as (4), and that these m-1 conditions determine the m-1 independent $\alpha_i = n_i/N$. Also, recall that the optimal $\alpha_i = n_i/N$ are independent of C. Then a solution to (4) for any value of C will yield the optimal α_i for (14). And, given the optimal fractional allocation $(\alpha_1,\ldots,\alpha_m)$ for (14), it is easy to find the optimal absolute allocation $(n_1,\ldots,n_m)$ and budget C which make $\phi = \phi_o$.

8. Additional constraints on the design problem (4) may often be required. The simplest sort of addition would be to replace some or all of the zero bounds $n_i \geq 0$ by positive bounds $n_i \geq n_i^0 > 0$. For example, the experimenter may start with observations numbering $(n_1^0,\ldots n_m^0)$ from some previous round of experimentation. This kind of additional constraint on (4) can be handled with only trivial modifications to the simple optimization conditions (7); so the design problem is still computationally straight-forward. More substantive additional constraints may also arise. For example, a fixed experimental capacity may add a total sample constraint $\Sigma_{i=1}^m n_i \leq N_0$. (Of course, the budget constraint may all along have been interpreted as a total sample constraint, where $c_1 = \ldots = c_m = 1$ and C is the maximum total sample. The current discussion refers to joint imposition of a budget and a total sample constraint.) Or a bureaucratic ruling may place a sub-budget constraint $\Sigma_{i=1}^p c_i n_i \leq C_p$ on the first $p < m$ design points. And so on. Such substantive additional constraints make the programming problem to be solved much more difficult. Nonetheless, if all the constraints are linear, efficient computer routines are available. Some types of non-linear constraints are, of course, also tractable.

III. THE APPLICATION

In the Graduated Work Incentive Experiment we are concerned with the response of family earnings to the changed alternatives produced by introduction of a negative income tax. Accordingly, we have specified as the dependent or response variable for a given family (the y_r in equation (1)) the ratio of the family's actual earnings during the experiment to a pre-experiment estimate of "normal" earnings. Calling the response variable R and dropping the subscript:

$$R = \frac{\text{actual earnings}}{\text{pre-experiment normal earnings}} .$$

The three independent or design variables specified were:

$$g = \frac{\text{maximum benefit (paid when earnings are zero)}}{\text{poverty level income}} .$$

$$t = \text{marginal tax rate} = \text{reduction in benefit per dollar earned}.$$

$$w = \frac{\text{pre-experiment normal earnings}}{\text{poverty level income}} .$$

So the response function was of the form $R = f(g,t,w)$ (neglecting the error term). The first two design variables g and t are subject to direct experimental control for each family in the sample; they are parameters of the linear negative tax the family is faced with. The third design variable w must be controlled by stratification; families must be screened until ones of desired w-level are found.

There are assuredly more variables than g, t, and w which may affect the response variable R; although many variables which come to mind may operate principally through g, t, and w. For example, family size operates through g and w by affecting poverty level income in the denominators of g and w. Also, family size, as well as education, race, age, and so on, operate through the normal earnings variable w. Insofar as variables excluded from f(g,t,w) can be randomized by careful sampling procedures, they can be lumped into the error term (the e_r of equation (1)). So the regression model used for sample design is a very abbreviated version of the model one might eventually apply to the data produced by the experiment. The practical limit on the number of variables that can be handled in the design is far more stringent than the practical limit on the number of variables that can be measured for eventual analysis. Even screening enough families to get the desired stratification by w-level turned out to be quite difficult. In summary, it seemed to us that w was clearly the most important stratification variable to control, and that once it was controlled there did not seem to be any second variable of comparable importance.

The problem, then, was one of specifying a sample in the three dimensional design space of (g,t,w)-triplets. Sampling was restricted to a region within the design space which provided substantial variation in (g,t,w), but which kept to (g,t,w)-combinations of actual policy interest. Within this region of interest, twenty seven design points were selected (so $m = 27$). There were nine (g,t)-combinations or treatments (one "control" combination with $g = t = 0$ and eight non-zero combinations) at each of three w-levels. So the design problem reduced to finding optimal numbers $n_1,\ldots,n_{27}$ of families to allocate to each design point.

A crucial part of defining optimality of a design is specification of a regression functional form. In the Graduated Work Incentive Experiment, numerous alternative transforms and combinations of g, t, and w were used to provide f(g,t,w) functions which had both linearity-in-parameters and varying degrees of non-linear flexibility in g, t, and w. The functional forms used had from 6 to 13 parameters; so a substantial degree of non-linearity in g, t, and w was allowed for.

For a given regression functional form, the variance matrix $V(b) = \sigma^2(\Sigma_i n_i x_i' x_i)^{-1}$ of the parameter estimates is easily obtained for any specific allocation of families to design points (that is, any choice of $n_1,\ldots,n_{27}$). The remaining problem in defining optimality of design lies in specifying a scalar-valued function of V(b) to optimize. In terms of the optimand $\phi(n_1,\ldots,n_m) = \sigma^{-2}tr[P'WPV(b)]$ introduced above, this requires specification of the matrices P and W. Recall that $P\beta$ is the assumed vector of magnitudes-to-be-predicted and thus Pb is the estimate of interest.

In the Graduated Work Incentive Experiment, the assumed objective was taken to be estimation of the incremental treasury cost of a linear negative tax due to induced reduction of work effort and earnings--that is, the difference between the cost assuming zero work reduction and the cost given the actual work reduction. (The possibility of negative work reduction--work increase--is fully allowed for.) The cost referred to is the cost for the entire country of a national negative tax. Given an estimate of the response function parameters, the work response and thus the desired incremental cost can be estimated for any individual family. By summing over all families, the national cost can be estimated. Such a cost estimate depends on the specific negative tax parameters g and t assumed as well as on the parameter estimate b; so the cost might be denoted H(g,t,b). (The variable w has "integrated out" in the summation.) With a few approximative tricks, this estimate can be expressed as a linear function of b, call it h(g,t)b where h(g,t) is a row vector depending on g and t. The rows of P were set equal to the values of h(g,t) for various policy-relevant combinations of g and t. So the elements of the vector Pb are estimated incremental treasury costs due to induced earnings response for various (g,t)-combinations. A policy-importance weight was specified for each (g,t)-combination, and thus each element of Pb; these weights were arrayed in the diagonal matrix W. This completed the specification of the objective function $\phi(n_1,\ldots,n_m) = \sigma^{-2}\mathrm{tr}[P'WPV(b)]$.

This objective function is to be minimized by appropriate allocation of families to the 27 design points--that is, by appropriate choice of $n_1,\ldots,n_{27}$. But this allocation was constrained by several further considerations. The principal constraint was the budget constraint $\Sigma_i c_i n_i \leq C$. The costs c_i were composed of administrative costs, which are relatively constant over design points (costs of screening families, administering questionnaires, mailing checks, and so on), and of negative tax payment costs, which vary widely over design points. A curious feature of this experiment is that the cost of payments to families is both an ingredient of the design problem and in essence what the experiment is designed to estimate. Nevertheless, we have some well-founded notions about the relative sizes of these costs, which do not vary widely even when assumed earnings response patterns do vary widely.

In addition to the budget constraint, there were sample size constraints, both on the total sample and on sub-samples for various w-levels. There was also a constraint on the fraction of the budget that could be allocated to a particular high-benefit (g,t)-combination. Finally, most of the design points had positive rather than zero lower bounds ($n_i \geq n_i^o > 0$) since a subgroup of families had been allocated to the design points prior to the use of the optimization model.

Allowance was also made for attrition from the sample of families at low payment design points, and for higher error variances of families at high payment design points.

These specifications combine to make up a programming problem involving 27 variables, a non-linear objective function, and typically five substantive linear constraints in addition to the 27 lower bound constraints. A computer routine due to Kreuser (1968) solved this problem in about 10 to 15 minutes on a Burroughs 5500. Of course, the problem was solved many times over as specifications were polished and sensitivities to changed specifications were tested. In addition to solutions for this problem, a small number of solutions were generated for a larger problem. A distinction was made between two locations at which families were being sampled; so there were 27 design points for each of two locations, or 54 design points and corresponding n_i altogether. The number of constraints and the number of terms in the regression function were also increased for this enlarged problem. This raised the solution time on the Burroughs 5500 to the neighborhood of 45 minutes.

The final solution of the Graduated Work Incentive design had the following noteworthy characteristics. (i) The optimal allocation of experimental families produces a decidedly non-orthogonal design. (ii) Several of the design points were allocated no observations or the minimum possible observations (given the positive lower bounds). (iii) The majority of the budget was allocated to a few high-payment design points; and the majority of the total sample was allocated to low-payment design points. (iv) The optimal designs generated by the model were substantially more efficient than various intuitive designs discussed before the optimization model was used. (Since the objective function ϕ is a variance magnitude, the relative efficiency of two designs can be measured in the usual sense by taking the ratio of their ϕ-values.)

Summing up our experience with the model in this, its first, application, we have found it very useful for incorporating explicitly a great many considerations which affect an experimental sample allocation. Indeed, it is difficult to see how these matters could be properly reflected by any less formal (or rule of thumb) procedure. The basic model will be used in the design of several upcoming social experiments, and its generality is such that it can be adapted to many experimental situations which are characterized by several dimensions of experimental control, and where continuity of response suggests some form of estimated regression surface.

ACKNOWLEDGMENTS

The research reported here was supported by funds granted to the Institute for Research on Poverty at the University of Wisconsin pursuant to the provisions of the Economic Opportunity Act of 1964. Thanks for helpful comments on this material are due to numerous members of the Institute for Research on Poverty and of Mathematica Inc. (of Princeton, New Jersey), as well as to Richard Day and Arthur Goldberger. Thanks are also due to Felicity Skidmore for editorial help and to Jerome Kreuser for non-linear programming help.

REFERENCES

[1] G. E. P. Box and H. L. Lucas (1959), "Design of Experiments in Non-Linear Situations," *Biometrika*, Vol. 46, pp. 77-90.

[2] Arthur Goldgerger (1964), *Econometric Theory*, New York: Wiley, 1964.

[3] J. Kiefer (1959), "Optimum Experimental Designs," *Journal of the Royal Statistical Society Series B*, Vol. 21, pp. 273-319.

[4] Jerome Kreuser (1968), *Users and Programmers Manuals for GPM*: University of Wisconsin, Social Systems Research Institute Programming Series, 1968.

[5] K. D. Tocher (1952), *Biometrika*, Vol. 39.

A NEGATIVE TAX EXPERIMENT FOR RURAL AREAS*

D. Lee Bawden, University of Wisconsin

Increasing concern about the exclusivity, inequity, and inadequacy of the present welfare structure has stimulated a search for alternative income maintenance schemes to augment or replace current public assistance programs. One of these alternatives is the negative income tax. An unique social experiment financed by the Office of Economic Opportunity is now under way in New Jersey to test the workability and consequences of this program for families residing in urban areas. This paper reports the plans for a second experiment, the purpose of which is to measure the effects of alternative negative income tax (NIT) programs upon rural people.

The paper will be organized as follows: (1) rationale for experimentation, (2) rationale for an additional experiment in the rural sector, (3) a discussion of the objectives and design of the planned rural experiment, and (4) some comments on social experimentation.

Rationale for Experimentation

The desirability of adopting a nationwide negative income tax depends, among other things, upon: (1) the cost of the program and (2) its effect upon the behavior and attitudes of the poor. While some answers can be provided from our existing state of knowledge about these issues, others cannot.

Taking the current income distribution, the cost of any specific negative income tax plan can be quickly calculated, assuming no change in earned income of the recipients. Unfortunately there is little evidence to support or refute the assumption that individuals would not alter their work habits and, consequently, their earnings if a negative income tax were introduced. In fact, this issue is basic to the arguments advanced by opponents of the negative tax -- that the poor are by nature shiftless and if guaranteed a minimum income, regardless of the negative tax imposed, a good many of them would prefer to accept that guarantee in lieu of working. Obviously some evidence exists to allay the fears of the most pessimistic.[1] But there is little evidence concerning just how much work disincentive, if any, various NIT alternatives would induce. Obviously, the more work disincentive, the greater the program cost since about half the money income presently received by the poor is earned from labor. Experimentation seems desirable in order to gather more information with respect to this issue.

The cost of any negative income tax program must be balanced against the benefits, and these benefits will be mostly in terms of changes in the attitudes and behavior of the poor upon receiving transfer payments. Judgments can be made about some behavioral changes by studying the (unassisted) near-poor. An example might be expenditure patterns. However, the poor are considered by many, and perhaps rightly so, to be different from the near-poor -- they are poor <u>because they are different</u> or they are different <u>because they are poor</u>. In either case one would expect their behavior after being elevated above the poverty line to be different from those who, unaided, have been able to avoid poverty. A good example of this is work behavior. Some believe that the poor are that way because they are lazy; others believe that the poor have less inclination to work because of past failures to find or hold jobs. On the other hand, many would consider the near-poor to be reasonably industrious, for to have evaded poverty they must be holding part-time or full-time jobs. Another example is the behavior of children of the poor -- they may well react differently after receiving transfer payments than those who come from near-poor families.

Evidence is also lacking on the effect of the negative income tax on other behavioral and attitudinal characteristics of the poor. An experiment simulating one or more negative tax programs would provide substantial information about many of these issues. For this reason, the Office of Economic Opportunity is now conducting a negative income tax experiment among urban families in New Jersey.[2]

Reasons for a Rural NIT Experiment

The New Jersey experiment is expected to yield a great deal of information about the effect of various negative tax plans on attitudinal and behavioral characteristics of urban wage earners. There is reason to believe, however, that these results may not be directly applicable to the rural sector, in which one-third of the nation's poor resides. One expected difference between rural and urban residents is in their work response to such a program -- because of differences in alternative employment opportunities and in the proportion of self-employed people. An accurate estimate of the magnitude of disincentive, both rural and urban, is crucial to estimating the cost of a nationwide program.

A negative income tax is also expected to have a substantial effect on the rate and composition of migration, both intra- and inter-community. Net migration out of rural areas is expected to exceed 10 million people during the 1960's and gross migration may be double that amount. Since there is considerable interest by policy-makers in ways to reduce and/or direct this flow, it seems important to learn the effects of NIT payments upon rural-urban migration -- both for evaluating the likely effects of a comprehensive NIT program and for gathering information useful in designing specific programs to induce or retard migration.

Also, it is not readily apparent that the specific program most effective for addressing urban poverty problems is best suited for rural poverty. For example, a large number of rural residents with low incomes are operators of farms or businesses in small towns. Determination of annual income as well as the appropriate timing of payments for the self-employed are likely to be different than for wage earners. This would be especially true for those farmers who receive their entire annual income at harvest time. The provisions for self-employed individuals in the New Jersey experiment are admittedly simple and probably inadequate for a nationwide, comprehensive NIT program.

The New Jersey experiment restricts eligibility to families of two or more members, with a male head between the ages of 18 and 58. Since a large number of poor households are headed by a female of working age, a study of their work behavior is also necessary for an accurate estimate of the cost of a nationwide NIT program. There seems to be no obvious way to infer from male work behavior the effect of this program upon female heads.

The second major category excluded from the New Jersey experiment is the aged. Men and women over 65 years of age make up one-sixth of the poor people and head about one-third of the poor households in the United States. While the work incentive issue is of less significance for this age group, it is nevertheless important. Also, of interest is the effect of a negative income tax on migration, spending patterns, and attitudinal changes of older people.

This need for experimentation in addition to that being conducted in New Jersey was recognized by the Ford Foundation, and a grant was made to the Institute for Research on Poverty at the University of Wisconsin to plan for a rural experiment. Such an experiment has been designed, and the first stage has been funded by the Office of Economic Opportunity. This experiment will be discussed in some detail below.

The Dispersed Rural NIT Experiment

Under the Ford grant ten staff members at the University of Wisconsin, all affiliated with the Institute for Research on Poverty and representing the disciplines of economics, agricultural economics, sociology, political science, law, and social work, combined in an interdisciplinary effort to design the rural experiment. It involves the selection of a dispersed sample of about 800 rural (farm and rural nonfarm) families; 600 of these will be headed by a male aged 18 to 58, 100 by a female in the same age range and 100 headed by a male or female over 58 years of age. The sample is being drawn now, and the first payments will be made this fall and will continue for a three-year period. The estimated total cost is $3.3 million.

The experiment is patterned after the one in New Jersey: it has the same basic objectives, a comparable experimental design, similar accounting periods for determining income and making payments, and it will be of identical duration. It differs from the urban experiment in that eligibility will be extended to single households as well as those headed by females and the aged, and a third accounting plan will be experimented with. Minor variations also exist in the definition of earned income.

Each of the major facets of this experiment will be briefly discussed below.

Objectives of the Experiment

The primary objective is to measure the effect of alternative tax rates and minimum guarantees upon the work incentive of rural residents and to compare and contrast these findings with those of New Jersey. This issue remains of paramount importance because a major hurdle to adoption of a nationwide NIT program is the commonly-held belief that payments, even with the negative tax, will significantly reduce the work effort of able-bodied males.

The second principal objective is measuring the effect of alternative NIT plans on the rate and composition of migration, with particular attention given to differences in response among age groups.

Of secondary importance are a host of other objectives, one of which is to learn the effect of payments upon the children of the poor -- their health, school performance, peer and reference group involvements, attitudes towards authority, delinquency rates, vocational aspirations, and numerous other characteristics. Changes in expenditure patterns are also of interest -- the distribution among savings, investment, and consumption; relative expenditures on necessities and luxuries; marginal expenditures on medical and dental care; and the effect upon credit vs. cash buying. Other objectives include the effect of NIT payments upon adult education (including job training), family structure (separation and divorce rates), involvement in social, business, and political organizations, family health, and attitudes towards one's self and others.

Location

The sample is being drawn from two separate locations, one in the South (North Carolina), the other in the Midwest (Iowa). The alternative of taking a nationwide rural sample was rejected in deference to administrative ease and a smaller operating budget. The choice of two areas rather than one is made because policymakers may distinguish between northern and southern rural residents. By selecting two locations, regional and ethnic differences in work incentive, migration, and other behavioral characteristics can be tested. The South is chosen because it contains a higher incidence of rural poverty than any other area in the United States. The Midwest is selected because it is (as classified by the USDA) "a relatively affluent area with a poor white minority." Since the Midwest itself is not particularly depressed, there is not a high degree of unemployment (desirable condition for an experiment designed to measure work disincentive.)

Criteria for selection of the specific counties in each region include the size and number of rural towns, their proximity to large cities, density of the farm population, diversity of agriculture, and representativeness of the entire region with respect to incidence of poverty, unemployment, racial mix, age distribution and educational level.

Experimental Design

The experimental design is similar to that in New Jersey. Families are being selected randomly from predesignated areas and, upon acceptance into the program, will be randomly assigned to a control group or to one of the program alternatives. Individuals will remain on that assignment for the duration of the experiment, and will be eligible for payments for the 36-month period regardless of their subsequent geographic location, as long as it is within the United States.

Family income at the time of screening must not exceed one and one-half times the established poverty line. These poverty levels are shown below for various family sizes. The poverty level will be adjusted annually to account for increases in the national cost of living.

Size of Household	Full Poverty Levels	
	Marginal	Total
	dollars per year	
Household Head	1688	1688
Spouse	791	2479
First dependent	528	3007
Second dependent	475	3482
Third dependent	422	3904
Fourth dependent	369	4273
Fifth dependent	316	4589
Sixth dependent	264	4853
Seventh dependent	211	5064
Eighth dependent	158	5222
Additional dependents	0	5222
Other adults	950	--

Any household which can establish eligibility for public assistance (including aid to the permanently and totally disabled, old-age assistance, aid to the blind, aid to families with dependent children, and general assistance) must choose either welfare or NIT payments, but not both. NIT payments will cease as soon as the household receives welfare benefits and resume according to the regular schedule as soon as such payments stop.

To insure a wider variation in environment the sample density will be fairly sparse, but not so much as to make selected individuals oddities in their communities. The sample will be stratified according to income level.

Program Alternatives

Five program alternatives involving three tax rates and three guaranteed minimums, will be tested. These are shown below for a family of four.

Poverty Levels	Tax Rates		
	30%	50%	70%
	Guarantee Level/Cut-off		
1/2		1741/3482	
3/4	2611/8703	2611/5222	2611/3730
Full		3482/6964	

About 50 percent of the families will be assigned to a control group and 50 percent assigned to the various plans, probably with less proportionate sampling in the more expensive plans.

Definition of Income

Income will be defined as the total gross income in cash or kind received by the household from all sources (including social security payments, unemployment compensation, strike benefits, and veterans' benefits). In some cases an imputed income will be added to reported income, with payments based on the total. Farmers' reported income will be increased five percent to reflect the value of livestock and livestock products (milk and eggs) produced and consumed on the farm. (Garden produce consumed at home will be ignored for the farmer and nonfarmer alike). Homeowners will have an imputed rental value added to their income. Finally, a percentage of net capital wealth will be added to income to reflect both earnings and potential capital consumption, the latter reflecting the thesis that the poor should, in part, "live off their assets rather than the Government."[3] The first $20,000 of business assets and the first $10,000 equity in owner-occupied homes will be excluded from net capital wealth for purposes of this imputation.

Payment Interval and Income Accounting

Payments will be based on income and the number of dependents as reported on returns filed by the participants. These returns will be filed every four weeks, showing gross receipts (wages for salaried employees, cash sales for businessmen) and (for the latter) cash expenses. Businessmen will report depreciation and other non-cash costs once a year, after filing their positive tax returns. All households will be paid biweekly, but the accounting period for computing income, upon which those payments are based, will be treated as an experimental variable.

One extreme is an income accounting plan that attempts to respond to a current lack of income -- to "fill the gap" so that a household's income does not fall below some predetermined level (this is the intent of most existing welfare programs). Obviously this is an impossible goal to achieve unless a family can accurately forecast its future income; payments based on a reported income for some past period cannot, by definition, be responsive to current income needs unless that income never varies. This objective can be approximated, however, by basing payments only on income for the preceding four-week period, the assumption being that if a wage earner loses his job, his family can live for four weeks on the wages paid for the preceding four weeks' work. At the end of the four weeks of unemployment, the NIT payments would respond to this situation, providing money for living expenses for the subsequent period.

This is the rationale for one of the income accounting plans to be used in the rural experiment -- the one-period (four-week) accounting plan.[4]

A second income accounting plan to be experimented with is the 13-(four-week) period moving average, i.e. an average of the preceding 52 weeks' income. In contrast to the one-period plan, the 13-period moving average is quite unresponsive to current needs resulting from fluctuating income levels. It is more appropriately viewed as a stabilized income supplement, most relevant to those with a fairly steady, but chronically deficient, income. As in the first plan, income is reported every fourth week and NIT payments are made biweekly.

The third accounting plan to be used lies between these two extremes -- it is a three-period moving average, with each period representing four weeks as below. The majority of households in both the rural and urban experiments will be under this plan.

The experimental significance of testing three accounting plans is not in their varying responses to income needs, but rather in the related issue of varying response of household behavior to the different plans. A primary objective of the experiment is to measure the work response of households to alternative negative tax rates. But the tax rates are not independent of the accounting period. Under the one-period accounting plan, the tax bite of an increase in work effort (and hence earned income) is felt immediately and fully in the next two biweekly NIT checks. Under the more retrospective 13-period plan, only one-thirteenth of the tax bite is felt in the following two biweekly NIT payments, i.e. the immediate effect of a change in work effort on NIT payments is diluted. But, by the same token, that effect is felt over a longer period--a full year under this particular plan.

This phenomenon can perhaps be more vividly illustrated by an example in the reverse direction. A person under a one-period accounting plan might be more tempted to remain idle for a month, knowing that his next two NIT checks will guarantee him a minimum income for that period, than if he were under a 13-period plan in which only one-thirteenth of that minimum guarantee would be reflected in the next two checks.

At issue, then, is a person's perception of the consequences (in terms of NIT payments) of a change in his work behavior, which in turn is partially dependent on the length of his planning horizon and on his assumed discount rate of future earnings.

Measurement and Analysis

The experimental households will be interviewed quarterly to gather information on the previously mentioned attitudinal and behavioral characteristics. Information will also be gathered from sources other than the families, such as schools, employers, hospitals, public organizations offering services to the poor, and other relevant institutions and organizations.

Some Comments on Social Experimentation

Experiments designed to measure physical responses of both human and nonhuman agents are common, especially in the area of medical science. Psychologists have performed social experiments involving animals. There have also been social experiments involving human beings, but these fall principally in the areas of business games, consumer panels, and observed group interaction. Social experimentation of the nature and magnitude of the negative income tax experiments is unique. Some possible problems can be foreseen; others cannot.

The Hawthorne Effect is obviously a matter of concern. Since families involved in these negative income tax experiments are generally unaware of what we are trying to measure, let alone what responses we expect, we are hopeful that this will not lead to a serious bias.

The greater problem is contamination of the experiment by the communication media. There has already been considerable pressure by representatives of television, radio, magazines, and newspapers to learn the details of the New Jersey experiment and to interview the families involved. Obviously, this type of publicity could easily lead to serious experimental bias. To date, the communications people have acted quite responsibly when explained the implications of their actions, but it is not clear that these arguments can successfully stave off their efforts for the entire duration of the experiment.

Local changes in the employment situation or the wage structure will have an influence upon the measured responses of individual participants. This is one of the reasons that both the urban and rural samples are not only dispersed within a specific area, but are drawn from more than one area. This, however, does not control for external national changes such as in the general level of unemployment, fiscal spending, or inflation.

Another potential problem in this type of social experimentation is a change in the laws governing the actions of the participants, or their benefits or liabilities to society. For example, if the Family Assistance Program is enacted as proposed with a guarantee of one-half the poverty level, it will be roughly comparable to the lowest plan in the urban and rural experiment and thus will have little or no effect on the experimental households receiving NIT payments. However, it will change the status of many of the control families, causing some difficulty in the comparative analysis.

Finally, there is just the problem of keeping track of the original sample. In the first four months of the New Jersey experiment an attrition rate of five to six percent was experienced. Some of this was due to families moving and leaving no forwarding address. Obviously, this could potentially bias the experiment since these families may well differ in their response to the program from those who remain in the experiment.

There are other serious problems in this type of social experimentation, as well. A period of three years cannot hope to simulate a nationwide, comprehensive, long-term negative income tax program. Furthermore, the information available to participants in a nationwide program will be different than that available to those in the experiment. In a nationwide program there would also be considerable interaction among participants, among non-participants with respect to the program, and between participants and non-participants, all of which would likely influence the attitudes and behavior of those receiving transfer payments.[5]

In summary, those involved in these experiments recognize a good many limitations to an experimental approach of evaluating alternative policies. It remains to be seen whether the information gathered from these and similar efforts will justify the sizeable cost of obtaining it.

Footnotes

*Helpful comments were received from Harold Watts on an earlier draft of this paper. Others contributing to the project, and hence to this paper, include Bert Adams, Joel Handler, Joseph Heffernan, Robinson Hollister, William Klein, Robert Lampman, Charles Metcalf, Charles Meyer, and William Saupe.

[1]For example, it is estimated that of all the male heads of poor families in 1967 who were less than 65 years of age, 60 percent held a full-time job the entire year, 35 percent worked part of the year or had a part-time job all year, and only five percent did not work at all [1, p. 11].

[2]For a detailed discussion of the urban experiment, see Watts [3].

[3]For a discussion of the various forms that this may take, see Weisbrod and Hansen, [4].

[4]This plan, as well as the three-period accounting plan to be described, embodies a "carryover" provision. Earned income in excess of the breakeven, or cut-off, level is carried forward for a maximum of one year and is added to income in any period in which such income falls below the breakeven level. NIT payments are based on earned income plus any amount assigned to that period from the carryover.

[5]For an in-depth consideration of the problems of social experimentation, especially regarding income maintenance, see Orcutt and Orcutt [2].

References

1. Cohen, Wilbur J. "A Ten-Point Program to Abolish Poverty," Social Security Bulletin, 31:3-13, December, 1968.

2. Orcutt, Guy H. and Alice G. Orcutt. "Incentive and Disincentive Experimentation for Income Maintenance Policy Purposes," Am. Econ. Rev. 58:754-772, September, 1968.

3. Watts, Harold W. "Graduated Work Incentives: An Experiment in Negative Taxation," Am. Econ. Rev. 59:463-472, May, 1969.

4. Weisbrod, B.A., and W. L. Hansen, "An Income - Net Worth Approach to Measuring Economic Welfare," Am. Econ. Rev. 58: 1315-1329, December, 1968.

STRATEGY FOR A BROAD PROGRAM OF EXPERIMENTATION IN INCOME MAINTENANCE

Larry L. Orr,[1] University of Wisconsin

I. Introduction

The initiation of the graduated work incentives experiment being conducted by the Institute for Research on Poverty and MATHEMATICA in New Jersey and Pennsylvania[2] has raised the prospect (some might say the specter) of an ambitious program of experimentation in social programs in general, and income maintenance programs in particular. Indeed, the enthusiasm for this relatively novel technique in some quarters threatens, at times, to outrun the capabilities of the embrionic reservoir of experience and expertise in this largely untried methodology. In response to such pressures and in the interest of rational policy formulations in the area of income maintenance, a group of researchers at the Institute for Research on Poverty of the University of Wisconsin have, at the request of the U.S. Department of Health, Education and Welfare, devoted the last few months to the development of an overall research strategy in this area. Its scope of research has encompassed a wide range of program alternatives, program effects, and research methodologies. The discussion in this paper draws heavily upon the work of this group, although it is in no sense a comprehensive report of the group's activities; in particular, I have limited my remarks to consideration of income maintenance experimentation, although an important part of our work for H.E.W. has involved the delineation of non-experimental research strategies.

It is important to understand from the outset exactly what "experimentation" means in this context, and to distinguish experimentation from the related, but distinctly different, concept of "demonstration." An experiment attempts, through the exogenous manipulation of the environment facing various economic, social, or political decision-making units, to measure their behavioral responses to variations in a particular program or program feature. Viewed in terms of a multiple regression model, the experiment seeks to generate data for the estimation of the response (dependent) variable as a function of a number of independent variables, some of which are policy parameters which can be manipulated experimentally. To achieve this goal, the experiment must obviously include several different "treatments," (at a minimum, one "experimental" treatment and a "control" or "status quo" treatment), in order to obtain estimates of differential responses. For example, the New Jersey experiment includes nine distinct treatments (including the control group), defined by different combinations of the income guarantee and the "special tax rate" under a negative income tax, with the objective of estimating the earned income response surface over the guarantee-tax rate plane. The importance of such experimental variation is that it yields information about behavioral responses to a variety of possible program variations, both those included in the experiment and, by interpolation or extrapolation, others not included.

By contrast, what I shall call "demonstrations" typically involve little or no experimental variation of policy parameters. A uniform treatment is applied to a specified group or geographic area, often without even any attempt to define a comparable control group. Thus, it is difficult to rigorously test hypotheses in a demonstration; at best, one gets a qualitative feel for the consequences of the single program variant applied, and some idea of the administrative feasibility of the program.

While demonstrations of this type may be useful for certain purposes, in this paper I shall confine my attention to experiments in which hypotheses relating to behavioral responses to specified policy parameters can be rigorously posed and tested.

II. Criteria for the Selection and Design of Experiments

Experimentation in income maintenance is an extremely expensive research undertaking, not only in financial terms, but in terms of research talent, which at the moment is a very scarce resource in this area. It follows that experiments should be used sparingly and be carefully designed to maximize their informational output. A prime criterion for the selection of a particular hypothesis for experimental testing, then, is whether that hypothesis can be adequately tested by non-experimental (and, therefore, generally less expensive) methods. If relevant non-experimental data exist, the presumption is against using the experimentation to generate new data. The other side of this coin is, of course, that even if non-experimental data is not available, experimentation is feasible only if it can be reasonably expected to provide a definitive test of the hypothesis in question.

For those questions where both of these conditions are satisfied, one is faced with the problem of assigning research priorities which rank experimentable hypotheses according to some set of criteria. I would suggest that the following criteria be applied, roughly in the order presented.

1) Policy relevance. The over-riding objective of research on the field of income maintenance is to provide guidance to policy makers in the revision and modification of income maintenance programs. Therefore, an obvious and appropriate criterion in developing experimental research is the usefulness of the information to be obtained from such research as an input into the policy decision-making process. This does not mean catering to political whims or pressures. It simply means that certain

behavioral responses will bear more heavily upon the desirability of any particular income maintenance plan than will other responses; ceteris paribus, these responses should receive higher experimental priority. The focus of the New Jersey experiment upon work effort response is a case in point; clearly, the response of recipients' earned income will have a major impact upon the cost of a negative income tax, as well as its political acceptability in terms of the dominant Puritan ethic. Moreover, because this response might be expected to be most severe among families with male heads in their working years, the New Jersey experiment was limited to that population.

2) Replicability. As a second criterion, I suggest what Hollister and Cain have called the "replicability criterion."[3] This criterion would restrict experimentation to those program features which can feasibly be replicated on a national scale. Any number of programs can be devised and instituted on an experimental basis which, because of their cost or administrative complexity, could not reasonably be considered to be feasible alternatives for national policy. For example, one might hypothesize that a very intensive job training and counseling program would be an effective offset to the work incentives of cash transfers. It may well be, however, that even if such a program could be carried out experimentally, it would simply not be feasible to provide such services to all recipients of a national transfer program. A corollary of the replicability criterion is that to obtain valid estimates of the effects of a national program, it must be possible to replicate the hypothesized feature of the national program in the experimental setting.

3) Adequacy of existing theory and measurement techniques. In certain areas, the current state of the art severely circumscribes the possibilities for experimentation. In many cases, we have a vague notion that a particular policy parameter may have important behavioral effects, but we have only a general idea of the mechanism by which the effect will operate, and possibly only very crude quantitative measures of the effect itself. For example, one such area is the whole question of the effects of income maintenance on the institutional structure of the community. It seems plausible that large-scale income transfers will have important effects upon the interactions of recipients and nonrecipients within a whole range of economic, social, and political institutions. Yet we really have no coherent theory of community which would allow rigorous formulation of hypotheses for experimental testing of these effects. We don't know which policy parameters would be crucial to any given effects, and therefore should be varied experimentally, and--perhaps more importantly--we don't know which non-policy variables must be controlled for in order to assure valid inference.

By contrast, underlying the investigation of work effort response in the New Jersey experiment is a well-developed body of economic theory relating labor supply to wage rates, income, and other family characteristics. Where even the rudiments of such a theory are lacking, the wisest strategy is probably to devote our research effort to develop the basic theory before proceeding to experimentation.

Closely related to the theoretical underpinnings of experimentation is our ability to measure behavioral responses quantitatively. Our ability to pose meaningful, testable hypotheses, and to generalize experimental results is severely constrained by the sophistication with which we can measure responses. Returning to the example of community effects, we face the difficult problem of defining and measuring institutional change. Can we really define appropriate indices of political participation and tension, social adjustment or alienation, or adequacy of social services? If not, the results of experimentation are likely to be ambiguous at best.

The requirement of an adequate theoretical and measurement capability is especially important for the specification of an efficient sample design. The question of adequate sample size will be discussed later in this section; at this point, suffice it to say that efficient determination of sample size depends upon our ability to predict the "normal" variability of the response variable, given the values of relevant non-experimental variables (which must be specified by our theory or empirical data.) The sample must be sufficiently large to distinguish the impact of the experimental variables from this residual noise in the response variable.

The criteria just presented bear upon the selection of hypotheses for experimental testing. Once it has been decided that a particular hypothesis requires, and is amenable to, experimentation, the question of optimal experimental design arises. This is a complex question, and this is not the place for a detailed discussion of the statistical intricacies involved. However, it seems useful to establish certain guidelines for design which are relevant for the development of a comprehensive research strategy involving a number of separate experiments; the following include some of the more important considerations.

1) Experimental objectives. While the information obtained from any one experiment may be useful in analyzing a wide variety of questions, it seems most efficient to focus each experiment upon a single dominant response variable. This limitation is imposed by the necessity of defining a single experimental objective function which is to be maximized through the sample design. Maximization of the objective function is roughly equivalent to minimizing the error of estimate in the predicted response variable; it is, essentially, the efficiency criterion for the sample design. If an experiment attempts to focus on more than one objective, it is not at all clear what the criterion of efficiency in response estimation

should be.

This should not be construed to rule out the collection of data relating to a wide range of behavioral responses in each experiment. Indeed, one of the great virtues of experimentation is that it provides a rich source of longitudinal survey data on low-income households. Still, it seems most efficient to focus upon a single over-riding objective in setting the sample design. For example, the design model for the New Jersey experiment is based on the estimation of work effort response, even though the quarterly interviews are designed to elicit additional information on a wide variety of attitudes and behavior, ranging from family expenditures to political participation and social integration.

Selection of a single objective variable for each experiment has the additional advantage of permitting the selection of a relatively homogeneous sample population comprised of those households which seem most appropriate to the hypothesis being tested or which are most relevant to policy considerations. Homogeneity of the sample is desirable on several grounds. First, given financial constraints on sample size, it seems wisest to concentrate on that type of family for which a significant response seems most likely and/or important on policy grounds. Second, in many cases, it is not clear that a single functional form of the response function would be appropriate to the behavior of diverse family types.

Finally, concentration on a single objective allows the duration of the experiment to be tailored to the particular response variable under investigation. Most of the experiments currently underway or being contemplated entail payments over three to five years. A time horizon of this length is probably quite sufficient for the investigation of, say, work effort, which may be expected to respond to fairly short-term changes in income and wage rates. Other behavioral responses, however, may be determined by much longer-run income concepts; the retirement decision of older workers and family planning decisions are cases in point. To obtain valid estimates of these responses, a much longer payment period is probably required.

2) Comparability among experiments. One of the greatest potential values to be secured from a coordinated, national approach to experimentation in income maintenance is the ability to ensure comparability of the data collected in the various individual experiments. As noted above, experimentation can provide a rich source of cross-sectional and longitudinal data on the poor--a group for which existing data is notably meager. To be of greatest value, however, it is important that data from each of the projects be gathered on a comparable basis, so that it can be pooled for analysis. In some cases, (e.g., in the measurement of attitudes, motivation, aspirations, and the like), this will simply mean that the same interview questions should be asked in each experiment. More importantly, however, it means that the basic economic concepts, program features, and administrative arrangements should be held constant over all experiments, unless there are explicit reasons to the contrary. In essence, I am suggesting that a uniform set of "rules of operation" (the experimental equivalent of the statute governing a national plan) should be applied to all experiments. These rules would cover such things as the definition of the family unit and family income, filing and administrative procedures, and the timing of payments. Uniformity of operating rules would not, of course, preclude variation either within or between experiments of those policy parameters, such as tax rates and income guarantees, whose effects are to be studied experimentally. Uniformity would simply control for unwanted variation in those program features which are not of experimental interest, but which might act to confound the experimental results.

It should be emphasized that any one of the rules of operation might be selected as an experimental policy parameter. For example, one might wish to study the effect of variations in administrative arrangements (filing requirements, handling of claims, agency-beneficiary contracts, etc.), or of variations in the frequency of payments. The point is, that any variation either within or between experiments should serve some well-defined experimental purpose.

3) Sample size. Income maintenance experiments are an extremely expensive undertaking, relative to traditional social science research techniques. In the New Jersey experiment, the payments to households in the experimental group are averaging over $1000 per year, and even the control group families must be compensated for their time and trouble. Obviously, there is a premium on selecting as small a sample as is consistent with reasonably accurate response estimation, in order to maximize the information obtained from limited research funds.[4] As indicated above, a priori theory and empirical knowledge are extremely important in determining the minimum required sample size.

The approach which has been developed for solving the sample size problem for the O.E.O.-Poverty Institute rural negative income tax project is essentially an analysis of variance framework. The analysis requires an a priori estimate of the "normal" variance of the response variable (i.e., the variance in the absence of the transfer program), given the values of other relevant characteristics of the response unit. For example, if the response variable is family earned income, one would want to control such family attributes as education, occupation, family composition, etc., in estimating the year-to-year variance in family earned income.

Suppose now we wish to ask the question of whether the transfer has any significant effect on the response variable. One way of posing this question is to ask whether the difference between the mean response of the experimental group and the mean response of the control group is significantly different from zero, at some specified confidence level. For any given size of control and experimental groups, one can compute the "normal" variance of this difference and estimate the range of the response variable which falls within the specified confidence interval--i.e., the minimum response differential which can be detected with control and experimental samples of this size. If, for example, the standard error of family income, given family characteristics, is $600 per year, the standard error of the difference between control and experimental samples of 300 families each would be $69. This means that an observed difference of $136 per year would be significant at the 95% confidence level. Alternatively, this means that if average family income in the sample $4000 per year, a total sample of 600 families would be sufficient to detect a difference in earned income of about 3.4%.[5]

The prime requirement is, again, adequate a priori theory and empirical knowledge. The more accurately we are able to estimate the response variable in the absence of the experimental treatment, i.e., the smaller the residual error variance, the smaller will be the sample size required for any desired degree of estimation precision. Second, given the best available estimates of the normal variation of the response variable, a decision must be made as to the precision with which we wish to estimate the response. This latter decision is obviously conditioned upon the importance of the response for policy formulation; the smaller the sensitivity of policy considerations (program cost, for example) to the response variable, the larger will be the minimum level of response detection which can be tolerated and, therefore, the smaller the required sample size.

III. Priorities for Experimentation

Taking as our starting point the OEO work incentive experiments in New Jersey and the rural areas of Iowa and North Carolina, a number of possibilities for further experimentation suggest themselves. In this section, I shall apply the selection and design criteria presented above to obtain a priority ranking of those hypotheses which seem most amenable to experimental research. Heading the list of experimental objectives are a variety of issues in the broad areas of work effort response and changes in family structure. The experimental possibilities in these areas seem well within our current capabilities. A lower priority is assigned to experimentation focussed on the effects of income maintenance on community structure because, although there are a number of important issues in this area, experimental resolution of these issues does not seem feasible at this time.

In discussing experimental priorities, it is important to define the major program features, or policy parameters, which characterize any income maintenance plan. The characteristics which I consider most basic include:

a) the income guarantee; i.e., the payment which a family would receive if it had no other income. In general, this payment may be adjusted for family size and composition, and the schedule of guarantee adjustments will be an important factor in assessing certain behavioral responses;

b) the implicit tax rate; i.e., the rate at which the basic guaranteed payment is reduced as family income from other sources rises;

c) the definition of the family unit in terms of who may be included as dependents, who must be included as dependents, and who may qualify as a head of household;

d) the definition of family income and the accounting period over which income is measured for purposes of determining current payments; and,

e) coordinate programs which do not involve cash transfers (e.g., in-kind transfers such as job training, day-care facilities, and social services).

It is felt that the policy parameters listed here are general enough to characterize nearly any of the income maintenance programs currently receiving serious considerations. In general, there are three basic types of programs which have been widely advocated: negative income tax plans, children's allowances, and various modifications of the existing categorical welfare programs. Each embodies a particular guarantee schedule and tax rate, defines the family unit and family income in a particular way, and may be coupled with various coordinate programs. Therefore, rather than focussing upon program types per se, it seems preferable to analyze behavioral responses to these more general policy parameters. Proceeding in this manner, the hypotheses which should receive highest experimental priority are as follows.

1) Work effort response. The crucial dependence of program cost and the possibilities for the eventual eradication of poverty through income maintenance make the work effort response of recipients a question of highest research priority on grounds of policy relevance. Moreover, the crucial policy parameters of any national program which may be expected to influence work effort (guarantee schedules, tax rates, and the income accounting period) are features which are readily amenable to replication in the experimental setting. Finally, the existing theory and empirical knowledge of labor supply provide a sound basis for the design of experiments in this area.

a) The labor supply of families headed by non-aged males would seem to be adequately covered in the existing O.E.O. experiments. Further experimentation should focus on the work effort of the two other principal types of poor families, those with female and aged heads. The rural experiment will, of course, include some female and aged heads, but these small subsamples should be augmented with further experimental observations, especially in urban areas. These experiments would be closely patterned after the New Jersey design, in terms of treatments, sample size, sample allocation, and duration of payments. Since the normal work effort of such families is likely to be lower than for male-headed households, however, the cost of these experiments may well be somewhat greater than in New Jersey.

b) The work effort response of the aged (and near-aged) raises a unique problem which may not be amenable to the kind of experiment developed in New Jersey. Since an income maintenance program of the negative income tax type would constitute an assured retirement income which would be available at any time, such a plan might have a significant effect upon the age of retirement, especially for relatively low-income workers. Unfortunately, payments over a period as short as three years are probably not a sufficient inducement to elicit a reliable measure of the retirement response. It may be necessary, therefore, to select a sample of older workers who would be guaranteed income maintenance payments over the rest of their lives, to obtain a valid estimate of the effect of a permanent national program. A preliminary analysis of the response could be made after a fairly short interval--say, three or four years--although the experiment would continue to yield useful data for a much longer period of time.

Such an experiment would obviously be relatively expensive. However, costs could be reduced by sampling heavily at earned income levels near the break-even point, so that substantial payments would be made only to those workers who actually do curtail their work effort significantly; this would be entirely consistent with the dominant policy interest, since the majority of the retired poor presumably had incomes above the poverty line before retirement. Moreover, at age 65, Social Security benefits could be offset dollar-for-dollar against the transfers. Thus, if the sample consisted of workers in the 55-60 age bracket, one might expect to make large payments in only, say six or seven years to each household.

c) A third type of experiment in the area of work effort response which should receive high priority is the investigation of the effects of alternative specifications of the accounting period over which income is measured for purposes of determining payments. While at first blush this may appear to be relatively unimportant administrative detail, on closer examination it turns out to be a crucial feature of the transfer plan. The specification of the income accounting period has important implications for horizontal equity among households receiving the same average income over long time periods; payment levels under any particular accounting scheme could vary greatly from family to family, depending on the time-form of their income streams. In addition, the speed of response of payments to changes in family income, and therefore to emergency needs, depends critically on the accounting period; if payments are based on a lagged average of past income, as is usually proposed, they will adjust more or less slowly to changed circumstances, depending upon the length of the lag.

More fundamentally, from the viewpoint of behavioral response, the length of the accounting period may affect the recipient's perception of the marginal tax rate and, thus, his work effort. If the worker bases his work effort upon the return to labor over a fairly short time period, then a short accounting period, with rapid adjustment of payments to changes in earned income may be perceived as involving a higher tax rate than a longer accounting period with slower adjustment, even when the statutory tax rate on earnings is the same. Indeed, if this is the case, there is a conflict between the goal of making payments respond rapidly to need and the goal of minimizing the work effort disincentive.

In addition, the definition of the accounting period may create important incentives for recipients to manipulate the timing of their income stream in order to maximize payments. A short accounting period, for example, may induce greater seasonality in work effort.

These behavioral hypotheses seem important enough to merit explicit experimental variation in accounting periods, although it may be possible to integrate this investigation into those experiments which focus on the more general work effort issue. The accounting period for the majority of the New Jersey sample is a moving average of the previous three months' income with payments adjusted monthly; in addition, a small subsample will have payments based upon a moving 12-month average. The rural project is currently expected to involve three subsamples with different accounting periods: a twelve month moving average similar to the second New Jersey plan, a three-month moving average, and a four-week average. Unlike the New Jersey plans, however, the latter two plans will involve a "carry-forward" of any income above the family's break-even point, to be counted as income in later periods. This was deemed necessary to take account of the extreme seasonality of income flows typical of rural--especially self-employed--households. These plans suggest the range of possibilities in defining the accounting period. A number of further variants are made possible by the alternative methods of treating seasonal work-related and business expenses, alternative "inventory rules" for handling carry-forwards, and alternative rules for applying the carry-forwards to future income periods.

The sample sizes for several of the accounting periods proposed for the New Jersey and rural experiments will probably be inadequate to give anything more than a rough indication of the impact of the accounting period specification on work effort. These and other variants should therefore be tested in other experiments, especially in urban areas.

d) A fourth type of experiment would focus on the interaction of income maintenance with manpower, job training, and other work-related programs. It has long been argued that income maintenance programs which seek to preserve work incentives should be accompanied by programs which act to enhance the employability of the poor. The recent Presidential address on welfare reform, which explicitly tied job training, employment counseling, and day-care services to income maintenance, illustrates the concern of policy-makers with this issue. The underlying hypothesis for experimentation in this area are important interactions between these two types of programs; i.e., that the r combined effect would be different from the simple additive effects of income maintenance and work-related programs taken separately. One might hypothesize that the existence of income maintenance with work incentive features increases the attractiveness of the job training, while the availability of job training or day-care serves to reduce the disincentive effects which remain in the income maintenance program.

An income maintenance experiment to test these hypotheses is already in the planning stage. It is proposed that a variety of job training and counseling services and day-care arrangements be made available to families receiving negative income tax payments, with their response to be compared to a group for whom manpower programs, but not income maintenance is available. Unfortunately, since the experimental site was selected for the wide range of manpower programs available, it will be difficult to define a control group which has access to neither type of program. However, comparison of the results with the data from New Jersey, where manpower programs are less generally available, should provide a relatively reliable control. This, of course, places a great premium on preserving comparability between the two experiments in the design of the income maintenance program. In the case of day-care facilities, which will be included in the proposed experiment, it may be feasible to define a control group which does not receive these services, since they will presumably be subsidized for the experimental group.

e) A fifth experimental possibility in the area of work effort is the replication of the New Jersey model with a dispersed nation-wide sample.[6] This would provide a check on the generality of the results obtained in New Jersey and other experiments, by drawing observations from a variety of different environments and labor markets. In particular, by including observations from areas with high unemployment rates, one might obtain information on how work effort under a national program would vary with the level of economic activity. Moreover, such an experiment would provide observations in communities which fall between the small towns of the rural experiment and the large industrial cities of New Jersey and other urban experiments currently being contemplated, in terms of population size.

This undertaking would present some dificult administrative problems in maintaining contact with a widely dispersed sample. This might be reduced by cluster-sampling in a number of carefully selected areas, and/or contracting with a private survey organization which already has a national sampling capability. In any case, given these problems, this experiment probably should receive somewhat lower priority than some of those proposed below relating to the effects of income maintenance on family structure. Nevertheless, it has the potential for important contributions to our empirical knowledge of work effort response.

2) Effects on family size and structure. Virtually all of the income maintenance programs now receiving serious consideration provide potentially significant incentives for changes in the basic family structure of the recipients. In terms of policy relevance, these effects may well rival the effects of the program on work effort in importance. I would propose, therefore, that these effects receive an experimental priority just below the investigation of work effort response. Fortunately, experimentation appears to be feasible for at least the more important potential effects on family structure; the relevant policy parameters are readily identifiable and (at least approximately) replicable in the experimental setting, the response variables are easily quantified, and there is a substantial body of a priori empirical information upon which to base the experimental design. The following behavioral responses seem to be the best candidates for experimentation in this area.

a) Fertility. To the extent that fertility is influenced by the level or uncertainty of family income, any income maintenance program will be likely to affect family size. Perhaps more importantly, any program which adjusts payments by family size creates financial incentives to bear children. In the extreme, plans which determine payments solely on the basis of family size, such as a children's allowance, or which limit payments to families with children, as would the President's proposed Family Assistance Program, would seem to create a maximum incentive for increased fertility. Even a universal negative income tax plan would create such incentives, since it is usually proposed that the income guarantee be adjusted for family size. The importance of the fertility response is highlighted by the recent Presidential address on birth control and the long-standing (but virtually unsubstantiated) criticism of AFDC on the grounds that it fosters illegitimacy.

At first blush, it would seem that existing data might be sufficient to answer this question. A number of countries have adopted children's allowance, some (such as France's) at very substantial benefit levels. The evidence from these "natural experiments" is, however, ambiguous at best. The resulting birth rate patterns are rendered virtually unintelligible by the absence of any meaningful control group. Hence, experimentation would seem to be called for.

Unfortunately, short-term experiments in this area are unlikely to produce valid inferences as to the effects of a permanent national program. On the one hand, payments over three or five years provide a much weaker incentive to increase family size than would a national program providing payments over the entire 18 years of a child's minority. On the other hand, experimental families might be induced to shorten the spacing of their children in order to qualify for payments during the course of the experiment. It would be impossible to analytically untangle these countervailing effects upon birth rates in the experimental group.

To avoid these analytical hazards, it would be necessary to guarantee payments over a much longer period. For example, payments might be made over a period of 15 to 20 years, with adjustments in payments for any children born within that time. Analysis of results could be made after four or five years, although the experiment would continue to yield useful data for many years. This would substantially reduce, if not eliminate, both of the biases of a short-term experiment just mentioned.

The primary response rate of interest would be the birth rate, since increased family size may be expected to result in close spacing of children. Since it is conceivable that an increase in completed family size might result with no change in spacing, however, one would also want to gather data on such indicators as desired and expected family size. These would allow prediction of completed family size in the first years of the experiment, before most of the families reach their ultimate size.

The payments themselves might be structured in one of several ways. The obvious approach would be to provide treatments consisting of various negative income tax plans, patterned after the New Jersey treatments, over the entire course of the experiment. Although this would provide a valid simulation of all the major features of a corresponding national plan, it would be a terribly expensive undertaking.[7] This approach would also raise difficult administrative problems, since the income of the recipients would have to be monitored over the entire course of the experiment. Moreover, current funding of the entire experiment would be complicated by the unpredictability of income streams, and therefore payments, over such a long time horizon.

A second approach, which would substantially reduce these problems, would involve simulating only the features of a national program most relevant to the central issue at hand. One could create the "price effect" implicit in a negative income tax with family size adjustments by simply extending a flat annual payment (again, for, say 15-20 years) to each child born during the experiment. This payment plan would be, in effect, a children's allowance for additional children born to the sample families. Although children's allowances and negative income taxation are very different in many respects, the "price" each places on additional children is very similar, at least over a wide range of family income. This may be readily seen in Fig. 1, which shows total family income under a negative income tax as a function of earned income.

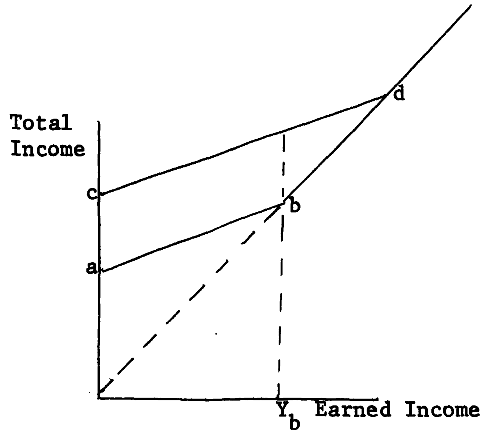


Figure 1. Total and Earned Income, Before and After an Additional Child.

Line ab is the schedule of total income (earned income plus transfers) up to the break-even income Y_b under the initial guarantee before the birth of a child. Line cd is the total income schedule after the birth of an additional child; the vertical distance between the two up to Y_b is equal to the guarantee adjustment for the child. Thus, the marginal increment to payments resulting from an increase in family size is a constant annual amount, regardless of the level of family income, <u>so long as earned income remains below the initial break-even level Y_b</u>. Over this income range, then, a negative income tax is indistinguishable, in terms of this price effect, from a children's allowance equal to the marginal guarantee.

Unfortunately, adoption of this payment scheme would preclude experimental analysis of the "income effect" of income maintenance on fertility, since payments would be generally lower than under a full-fledged negative income tax. I would argue, however, that this question could be adequately analyzed from existing data, and is therefore not worth the added experimental cost of a negative income tax.

The cost of payments of this type could be markedly less than under a long-term negative income tax. Suppose, for example, that a sample of families with two children were selected and guaranteed allowances of $400 per year for each additional child.[8] The present discounted cost of a 15-year allowance, then, would be about $3900, discounting at 6%. The average expected completed family size of families at these income levels is probably about 3.8 children;[9] let us therefore assume that on the average, each family has two children, spaced, say, one and three years after the beginning of the experiment. The present discounted cost per family would then be approximately $6,900.[10] This cost could be further reduced if the allowances were successively reduced for the fourth and subsequent children, as they would under a national plan.

A rough estimate of the required sample size for detection of a fairly small change in birth rates or desired family size at a high level of confidence can be made on the basis of existing data.[11] It appears that a sample of 300 experimental families would be sufficient to detect a difference of about 10% in the birth rate over a three-year period or 4% in desired family size, as compared to the mean of the control group, at a 96% confidence level. If this level of precision is acceptable, then the children's allowance type payments described here could be financed at approximately the same total transfer cost as the current three-year New Jersey negative income tax experiment.

b) A second aspect of family structure which merits experimentation is the question of the effect of income maintenance on marital stability, in terms of divorce, desertion, and separation. Several policy parameters may be of significance here. First, to the extent that marital instability stems from economic stresses within the family, the sheer effect of additional income, i.e., the income guarantee, may serve to reduce instability. It has also been widely suggested that the schedules of adjustment of the guarantee for family size and structure which have been proposed under most negative income tax plans will create incentives for family breakup.[12] Typically, a spouse receives a smaller marginal guarantee than the head of household. Thus, the family's total guarantee can be increased if the couple separates and forms two households, with two head of household guarantees. This incentive is analogous to the incentives for family breakup embodied in the AFDC program. Under AFDC, a family is eligible for payments only if the father is not present, and it is often alleged that this promotes desertion and separation. Of course, for that group of women currently on AFDC (or those who would become eligible at some time in the future), a change to a system which provides income maintenance for intact families would <u>reduce</u> the incentives for separation.

Perhaps more importantly, because of the discontinuity of tax rates at the break-even level of income, a family may be able to increase its total payments by forming two households, one with income above its break-even point and one with income below.[13] This kind of incentive would also be present in any plan which involved a non-linear tax rate, even if both of the new households remained below their break-even points. This effect is identical in principle to the incentive to file joint returns under the positive income tax, with its non-linear progressive rate structure.

We have only begun to consider the problems of sample design for such an experiment. Since nearly one-third of all divorces occur in the first four years of marriage,[14] it seems reasonable to select a sample of newlywed couples. It is also probable that economic considerations play a larger role in marital instability early in marriage than in later years.

While we have not undertaken a detailed analysis of the required sample size for such an experiment, a rough idea of the requisite sample size can be obtained if we assume that divorce is a stochastic binominial process with each family in the sample having an equal probability of divorce within the experimental period. If, for example, the incidence of divorce among couples of the type selected would be 20% in the first four years of marriage in the absence of income maintenance, a sample of 900 couples would be sufficient to detect a change of 2.7 percentage points in the divorce rate at the 96% confidence level.[15] In the relative terms, this is a rather wide confidence interval; it allows significant detection of changes no smaller than 13% of the initial divorce rate, with a fairly large sample. Even a sample of 2000 couples would allow detection of a relative change no smaller than about 9% of the normal divorce rate. Of course, estimation accuracy would be improved if one were able to predict the probability of divorce for individual couples more accurately than by simply applying the mean rate to all couples. Even so, it appears that a large sample will be required if we hope to detect even moderately small changes. It should be noted, though, that payments to couples with no children (or only one or two) under any given plan would be much less expensive than the average payment levels in New Jersey, where the families are fairly large. Thus, cost considerations are less restrictive with respect to sample size.

c) <u>Split-off of dependents.</u> Very similar to the incentives for marital instability are the incentives that may be provided for dependents to leave the family and set up new households. Again, the total family guarantee can be increased and payments may rise if the split alters the marginal tax rate facing either the original family or the new unit. The latter effect may be especially important if a dependent of a relatively well-to-do family can set up a new household with an income below the break-even point; this would be the case, for example, for teenage youths leaving families with incomes above break-even point. A variety of other dependents now living with families might also be encouraged to set up households of their own if afforded an income guarantee: grandparents, in-laws, unmarried relatives, etc. The fact that these potential income maintenance recipients currently reside in families throughout the income distribution means that the aggregate effect, in terms of program cost, could potentially be very significant. Of course, to the extent that these dependents continue to receive support from their families after leaving, this would count as income and reduce their payments.

While it would seem to be very desirable to attempt to estimate experimentally the impact of these incentives, the design of an experiment which would constitute a valid replication of a national program poses serious difficulties. The logical experimental approach would seem to be to select a sample of families to receive income maintenance payments, and to allow any individual who leaves a sample family, and any dependents he may acquire, to qualify for payments as a separate household. This approach, however, would overstate the incentives for splitting off embodied in the corresponding national plan, especially among young adults, because it creates a "dowry effect." In the experimental setting, a young man or woman would become a differentially attractive marriage partner because he or she would be eligible for income maintenance payments while his or her compatriots would not be. Under a national plan all single individuals would be equally eligible for income maintenance, so that all would compete for marriage partners on an equal footing. This asymmetry between experiment and national program then, creates an artificial incentive for others to "marry into" the sample which would tend to bias upward our estimates of the impact of the plan on marriage of dependents.

An alternative way of defining guarantees for individuals who leave the family to marry, which introduces a downward bias, might be incorporated into the experimental design in an attempt to bracket the true response under a national plan. This would involve allowing only the individual from the original sample family, but not his dependents, to receive payments; thus, again because of the asymmetry between experiment and national program, there would be an artificial scarcity of marriage partners eligible for payments, introducing a downward bias in rates of split-off. The dowry effect would still be present (in a weaker form), but if it is felt that this "scarcity" effect would outweigh the dowry effect, this subsample would provide a lower bound for the estimate of the true effect. The subsample where individuals are allowed to bring in new dependents would provide an upper bound.

For individuals or subfamilies leaving the family to live on their own, there is no problem with the acquisition of dependents. Thus, the experimental results in these cases would be relatively reliable.

We have not yet begun to explore the questions of sample size and design and program cost involved in such an experiment. Sufficient information for such an analysis is readily available, however, in existing cross-sectional data on the structure of families and subfamilies, and rates and ages of dependents leaving the family at various income levels.

3) Community effects. The experimental possibilities discussed so far relate to essentially individual responses of the income maintenance recipient. These can be estimated from the behavior of families in dispersed experimental samples with little or no interaction among recipients. One might also, however, expect an income maintenance program to have a variety of effects upon social, political, and economic institutions within the community as the result of interactions among recipients and between recipients and non-recipients, under a national program. To name only a few possibilities, there might be changes in the political power balance, governmental tax and expenditure patterns, attitudes toward the poor and aspirations of the poor, and location of economic activity and the economic opportunity structure facing the poor.

To attempt to induce these effects experimentally, one would have to adopt a strategy of "saturation sampling," giving income maintenance transfers to all households in a given area who would be eligible under a national plan. Moreover, the experimental area would have to be sufficiently large to encompass the geographical extent of the particular institution or activity being studied.

Such an experiment would obviously be very expensive. There are also several reasons to question the validity of the results that would be obtained. For one thing, as mentioned near the outset, we have very little in the way of a theory of community or of institutional change upon which to base the experimental design. Moreover, it is difficult to define measures of institutional change which would serve as response variables, or even to select the policy parameters which should be varied experimentally. More fundamentally, perhaps, an experiment in a single area, measuring the program's impact on one set of institutions, would, in essence, yield a sample of only a single observation. To constitute anything more than a demonstration (albeit, perhaps, a useful demonstration), experiments would have to be performed in a number of communities. Unfortunately, in the absence of an adequate theoretical foundation, it is impossible to say just how many community observations would be required for valid inference.

Given the design problems involved in saturation experiments and the range of other important questions more amenable to experimentation, it seems best to defer experimentation in this area, and to concentrate our research efforts instead upon developing the theoretical prerequisites to experimentation focussed on community effects.

IV. A Brief Preview of Coming Attractions

To date, two experiments in income maintenance have already been funded by the Office of Economic Opportunity--the New Jersey experiment and the rural experiment about to begin transfers in Iowa and North Carolina. Since these have been described here and elsewhere in some detail, I will simply note that in terms of the priorities proposed here, they focus primarily upon the work effort response of male headed families under a negative income tax. The rural experiment will also include some families with female and aged heads, but these samples will be small and the results not necessarily applicable to similar families in an urban setting. Finally, several variants of the accounting period will be tested in both experiments but, again, the samples are small and the alternative plans are not exhaustive. These experiments may generate some useful data with respect to other questions on my priority list, but they are not designed for this purpose, so such results will probably be only suggestive at best.

To expand our knowledge in those areas of high priority not covered by these projects, the Department of Health, Education and Welfare is now considering the initiation of a small number of carefully selected and controlled income maintenance experiments, and has already given planning grants for the design phase of two experiments. The focus of these two experiments will be upon issues which rank high on the list of priorities presented here.[16]

The first, located in Seattle, will consider the interaction of income maintenance with various manpower and work-related programs, along the lines suggested in section III.1.d. of this paper. The cash transfer plans will be closely patterned after the New Jersey treatments to ensure comparability of the data obtained. The sample will include a substantial proportion of female heads of household, however, so that we can begin to obtain estimates of the work effort response of females. The manpower programs to be associated with the transfer payments, for at least a subsample of the recipients, will include job training, employment counseling and referral services, and day-care facilities for female heads and wives with pre-school children. To avoid the costs and administrative problems of setting up new programs and facilities the recipients will simply be guaranteed a slot in existing programs and/or a subsidy to cover the costs of using the facilities.

The second experiment for which a planning grant has been made is to be carried out in Gary, Indiana. The experimental focus there will be on the work effort response under several variations in the income accounting period. Again, the sample will include a large proportion of female heads of household. This experiment will also conform closely to the form of the New Jersey experiment. It is hoped that the transfer phase of both the Seattle and Gary experiments can begin by the end of 1970.

H.E.W. has received requests for support of income maintenance experiments in a number of cities, and will undoubtedly solicit additional proposals for experimentation in several of the areas proposed here. The outlook for a fairly ambitious experimental attack on the large remaining areas of ignorance of the effects of income maintenance is quite promising.

Footnotes

[1]The research underlying this paper was supported by the Institute for Research on Poverty under a contract with the Social and Rehabilitation Services Division of the U.S. Department of Health, Education and Welfare. The author has benefited enormously from continuing discussions with a large number of members of the Poverty Institute; in particular, this paper draws heavily upon help from Robinson Hollister and Harold Watts. Needless to say, the author alone is responsible for the analysis and opinions advanced in the paper.

[2]For a detailed description of this project, see Harold W. Watts, "Graduated Work Incentives: An Experiment in Negative Taxation," American Economic Review Proceedings, May, 1969, pp. 463-472.

[3]Hollister, Robinson, and Glen Cain, "The Methodology of Evaluating Social Action Programs," Discussion Paper, Institute for Research on Poverty, University of Wisconsin, Madison, April, 1969.

[4]An alternative method of reducing costs is, of course, to concentrate on less generous transfer plans. Given our degree of ignorance about the magnitude and functional form of the response, however, it seems preferable to stick plans which seem generous enough to elicit a significant response.

[5]These calculations were made by D. Lee Bawden and Charles Metcalf of the Institute for Research on Poverty, University of Wisconsin, in preliminary design work for the O.E.O. rural negative income tax experiment.

[6]Such an experiment was originally advocated by Guy Orcutt, of the University of Wisconsin Department of Economics, now at the Urban Institute.

[7]If payments average $1000 per year, as in New Jersey, e.g., the present discounted cost per family for a 15-year program, discounting at 6%, would be about $9,700.

[8]Of course, experimentally, one would want to vary the payment among families, in order to estimate a continuous response function, but $400 ought to be a reasonable average guarantee. A negative income tax which guarantees the poverty line income would carry a marginal guarantee of about $400 for a third child.

[9]N.B. Ryder and C.F. Westoff, "Relationship Among Intended, Expected, Desired, and Ideal Family Size: United States, 1965," Population Research, March, 1969.

[10]In addition, it might be necessary to pay a flat annual allowance on the order of $150 to all families on the experiment to secure their cooperation in keeping in touch with the research organization during the early years when most families are receiving no payments. This payment could be terminated after four or five years when the analysis of results is undertaken, so that it would add only about $600 to the discounted cost. Such a payment would probably also be necessary under a negative income tax (in fact, this is being done in New J rsey), so it would not materially affect the comparison of costs under the two modes of payment.

[11]The calculations which follow were made by Glen Cain, Department of Economics and Institute for Research on Poverty, University of Wisconsin, on the basis of survey data reported in C. F. Westoff, R. G. Potter, P. C. Sagi, and E. G. Mishler, Family Growth in Metropolitan America, Princeton University Press, 1961.

[12]For a typical discussion, see "A Model Negative Income Tax Statute," Yale Law Journal, Vol. 78: 269, 1968, pp. 276-278.

[13]For example, a couple with a \$3000 income, a \$2000 guarantee, and a 50% tax rate would receive a transfer of \$500 if they live together. Suppose now they split up, with the husband retaining the \$3000 income and the wife none. If each is entitled to a \$1000 guarantee (to maintain the same total guarantee in this example), the husband would now receive no transfer, but the wife would receive \$1000, increasing their total payment by \$500.

[14]<u>Vital Statistics of the United States</u>: <u>1964</u>, Table 2-5, p. 2-8, Vol. III--Marriage and Divorce, U.S. Department of Health, Education, and Welfare, Public Health Service, Govt. Printing Office, Washington, 1968.

[15]This calculation was made in the following manner. The same variance is given by the formula:

$$\sigma^2 = \frac{p(1-p)}{N},$$

where p is the probability of divorce for any one couple and N is sample size. For N = 900 and p = 2, σ = .0134, and the 96% confidence interval is 2σ = .0268.

[16]This is not coincidental, since the objectives of these experiments were defined in consultation with members of the Poverty Institute where this list of priorities was developed.

DISCUSSION

James D. Smith, Pennsylvania State University

Bawden and Orr present a potpourri of interesting information about three social experiments. The Orr paper also ventures out into the area of meta-science and lays down some criteria for social science experimentation. The three criteria he specifies are:

1. Relevance
2. Replicability
3. Adequacy of theory and measurement techniques

The third of these is directly out of the researcher's primer and nothing more needs to be said about it. The first two, however, warrent some comment.

Unfortunately, relevance is not operationally defined. Every time I attempted to push hard on his criteria it came out sounding like: That which is relevant is that which I think is relevant. What Orr would like to do is lay down some guidelines to select research projects which have the greatest efficacy for changing peoples lives vis-a-vis the political decision making process. Presumably the most relevant projects would be those which changed the lives of the greatest number of people in the shortest period.

At one point he argues ".... an obvious and appropriate criterion in developing experimental research is the usefulness of the information to be obtained from the research as input to the policy decision-making process." He goes on to say that "this does not mean catering to political whims or pressures." But only a short distance on he cites the New Jersey experiment as an example of social research with the comment that "clearly the response of recipients' earned income will have a major impact on the cost of the negative income tax, as well as its political acceptability in terms of the dominant Puritan ethic." It seems to me that the selection of what to test for was quite sensitive to the political wind at the time the New Jersey experiment was being formulated.

The other criteria, replicability, seems to me to be a useful one for bureaucrats responsible for allocating research money and who have the potential intervention of the government into the society as a concern. There is no point in such people experimenting with what can be done for a small experimental group if there is little hope of replicating the effort for a much larger group, presumably national in scope.

Turning to the rural negative income tax experiment, I would first like to raise a few questions about measurement. It was mentioned that rental income would be imputed to home owners. In a panel study of income dynamics that has been going on at the Survey Research Center of the University of Michigan for the past two years it was found that between five and six percent of farm families get free housing in connection with their job. It might be desirable then to estimate the value of such housing in computing family income for the negative income tax experiment.

In the same panel study we found that just under five percent of family heads had severe work limitations caused by mental or physical conditions. And another approximate eight percent had some limitation on the amount of work they could do--such things as being unable to lift heavy weights, periods of pain and need for frequent rest periods. The incidence of work limitation is likely to be much higher among the poor than for a cross-section of the population. And, if one is interested in measuring work response to a money incentive, some account should be taken of the people who are precluded from responding by medical conditions.

One final comment on measurement. All of us who have been involved in data gathering have given lip service to the desirability of doing validation work along with field studies, but we have been grossly guilty of not doing it. In the case of the negative income tax experiment it would seem that the accuracy of reported financial information is of direct substantive importance. First, underreporting of income will result in overestimating the cost of such a program and, secondly, I would predict that one of the more articulated concerns of political decision makers will be the opportunity for cheating that a negative tax scheme will afford casual workers, the marginally self-employed and those in a position to substitute income in kind for cash flows. It also seems that these opportunities may exist relatively more in rural areas than in urban ones. I would strongly urge some validation work, at least on income reporting.

Finally, I would like to raise what seems to me to be a couple of rather knotty ethical questions and ask what you are doing about them. Since some of the families you will be experimenting with will have been on welfare programs which provide them with income in kind and social services, they will presumably be giving up their nonmoney benefits in exchange for cash. Cash will increase the consumption options of the family head but might have little effect on those of a wife and children. It is entirely possible that the dietary and medical needs of dependents could suffer with the change to a cash flow. Has any thought been given to the possibility that the experimenters may wish to intervene directly to alter an internal family mal-distribution of income? Another problem which is much more likely to occur involves the acquisition of consumer debt by people who have been assigned higher benefit levels in the experiment. When the experiment ends these people will presumably end up with a much smaller income. Since the society has used them as experimental subjects, a good case can be made for not leaving them any worse off than they were before the experiment.

DISCUSSION

Martin Frankel, University of Michigan

As the authors discovered, work in the area of optimal sample allocation for the estimation of what we might call "analytical" parameters is sorely needed. This lack of theory is compounded by the large gaps between the literature of experimental design and the literature of survey sampling. The experimental design and econometric literature usually assumes that we are taking simple random (with replacement) samples from hypothetical superpopulations. However, when the population of interest is widespread (e.g., families in the U.S.) practical economic probability sample designs are usually stratified and use clusters of sample elements. When regression models are used clustered samples usually induce positive covariance between error terms. Thus additional parametrization of the model would be required.

In the Graduate Work Incentive Experiment the formal population was a small area in New Jersey, and a systematic element sample was used. However, if it is deemed desirable to be able to make formal statistical inference to the entire U.S., a clustered sample would probably be used.

Independent of the type of probability sampling used (either simple random or clustered) there are two assumptions of the stated regression model which are probably always violated to some extent.

The violation of the assumption of homoscedasticity (equal error variances) is probably the less serious, resulting simply in a non-optimal allocation. However, assuming a known parametric form $y = bx$ may lead to serious problems when certain very disproportionate allocation schemes are used.

We are all aware that given enough real multivariate data we can usually reject any specific parametric hypothesis concerning the functional distribution of conditional means. This fact does not deter us from assuming a given form for the regression equation when we feel that the specific violations in various subranges of the design values will tend to average out over the entire range of interest. However, whenever most of the sample is allocated to a specific subrange of design values, the specific anomaly found in that subrange will be overrepresented in our estimate.

Thus in situations where the optimal allocation is extremely disproportionate, we might want to proceed with caution before we actually allocate the real sample.

DISCUSSION

Nelson McClung, Urban Institute

The Conlisk and Watts application of the principles of optimal allocations under risk to the design of social experiments is most helpful in assuring us that the activity is being conducted efficiently. Their analysis of research design is short one dimension--the timing of experiments. Clearly, a desirable feature of a good research design is completion of the experiment prior to adoption of the program whose effects are being evaluated.

Social experimentation is progress. It has been discussed idly for many years, always until quite recently rejected as unfeasible. But the next step must be to win acceptance for experimentation sufficiently prior to the time that there is general support for the programs that the results of experimentation can be used in program design. The Administration proposed vanishing income supplement for families with children is a case in point.

From a wish to avoid the unknowns of work incentive effects, the President has said that the supplement would be employment conditioned. Given that the Employment Services are State institutions and the responsibility for administration of manpower programs is to be transferred to State governments, the employment conditioning will be at quite local levels. Thus, there will continue to be opportunities as there are now in Public Assistance to use income supplementation as a device for maintaining an exploitable labor force. What sorts of sweated employments will become OJT programs?

Because the Family Assistance System is a program for a demographic category, it will have socially dysfunctional incentives not unlike those of AFDC, as people try to get into the favored category. The first child in a family of two adults is worth $1300 per annum. Whatever the program's impact in redistributing income, it should work wonders in the redistribution of children. No poor home can afford to be without one. But the program should promote birth control because no poor home can afford to have two.

There are other design features which we can anticipate will make trouble. Hopefully, Congressional consideration will improve the design of the program, although it too frequently has not in the past. So, we seem to be embarked upon social experimentation in the old fashioned sense of adopting a program with the thought that experience with it will suggest appropriate modifications. The difficulty with what we might call institutionalist experimentation is that experience alters the option set. Vested interests are created and expectations generated which effectively foreclose options which experience indicates would be superior except that they are inaccessible politically.

Assuming that an income supplement program essentially similar to that of the Administration is enacted next year, the need for income maintenance experiments will be not less but greater. Only the focus will shift. Experiments with higher guarantees and higher tax rates will be useful in redesign of the income supplement which we can anticipate. A number of experiments on technical features of the program, such an experiment on income computation periods, should be undertaken. Then there are experiments relating the income supplement to other programs. The Seattle Manpower programs experiment is an example. One of the more important of these is the relationship of health care to poverty and the need for income support.

The work incentive issue is far from settled and will not be settled until much more econometric work has been done on household labor supply functions. Experiments which identify age, sex, region, race, primary and secondary earners, primary occupation, and other possibly relevant dimensions of labor supply must be conducted, if we are ever to have reasonable accurate estimators of the labor supply response to transfers, taxes or grants.

IX

PROBLEMS OF ATTITUDE MEASUREMENT IN SAMPLE SURVEYS

Chairman, DANIEL LEVINE, U. S. Bureau of the Census

SELECTING THE QUESTIONS TO BE ASKED

Norman M. Bradburn, National Opinion Research Center and
Graduate School of Business of the University of Chicago

On her deathbed, Gertrude Stein is reported to have uttered these memorable words: "What is the answer? What is the answer? Ah, what is the question?" And that is precisely the state I am in when asked to talk about the problem of selecting questions for a questionnaire. What are the questions about questions that should be asked?

I would like to share with you today some of my reflections on the extremely vexing problem of selecting the questions to be used in a survey research questionnaire. My remarks are directed specifically at the problems of choosing questions relating to social-psychological variables, which are being increasingly used in survey work concerning social programs. I believe, however, that many of the considerations relevant to the selection of questions on social-psychological variables will be pertinent to the choice of questions on other types of variables.

My reflections have led me to focus upon the difficulties involved in formulating adequate decision rules for the selection of social-psychological questions. As I see it, there are five principal obstacles to the development of clear decision rules:

1. A lack of agreement among behavioral scientists about the appropriate social-psychological dependent variables that are relevant to particular social programs;

2. An inadequate conceptualization of those social-psychological variables that are suggested for study;

3. A relative lack of interest in systematic methodological research and survey measurement;

4. The relative underdevelopment of measurement theory in survey work as compared with the sophistication of sampling theory; and

5. The special historical and cultural problems that affect the phraseology of questions.

Let me discuss each of these obstacles briefly. The first two are intimately related to one another and concern the problem of knowing what concepts one wishes to measure. First of all, social psychologists themselves cannot agree on what the relevant variables should be, and second, because of the poor conceptual development of the variables that are suggested, there is not even a reasonable degree of consensus on their relative importance. Let me give as an example some problems that NORC faced in doing surveys connected with the evaluation of Manpower Training Programs. Increasingly, those involved in the evaluation of these programs believe that some kinds of social-psychological factors may be important in determining the effectiveness of training programs, and even in some instances, that particular types of social-psychological changes should be viewed as important outputs of the program in addition to increased skill level. The extension of interest to factors other than narrowly defined economic ones is to be applauded. It is lamentable, however, that behavioral scientists, who appear to have succeeded in convincing economists that there is something more to the world than economic variables, cannot now come forward with better suggestions regarding the social-psychological variables to be studied.

What are the kinds of variables that behavioral scientists might propose as relevant to job training programs? Some would suggest variables concerned with motivation, while others would emphasize the relationship between personal interests (or needs) and the characteristics of the job itself or the work environment; some might stress variables relating to job satisfaction, while others would stress variables relating to managerial or supervisory styles. Still others might view the important variables to be those related to an individual's general disposition toward society or alienation from work. In a general way these differing approaches can be grouped into those approaches concerned with individual differences and with variables that are conceptualized as being within the individual; those approaches that stress variables relating to work organization and the social setting within which jobs are performed; and those approaches that cover both fronts by pointing to the importance of the interaction between the two general types of variables.

It is easy to point out that there is a lack of consensus among behavioral scientists about the importance of particular variables. It is not so easy to explain why such a lack of consensus should exist, or to suggest things that could be done to improve the situation. Although I have no good evidence, I strongly suspect that the situation is fostered by two trends in the behavioral sciences that have been going on for some time. The first is that most research in the behavioral sciences limits itself to establishing the existence of relationships among variables, and gives practically no attention to the assessment of the magnitude of those relationships. If one reads the professional journals in these fields, one has the feeling that psychologists and sociologists worship a god of "statistical significance" and have an unquestioning faith that all things which are statistically significant are equal in importance in the world and that all things which are not statistically significant are totally unimportant. Closely allied with such a mystical belief and aiding it is the fact that the great preponderance of research is carried out under laboratory or other artificial conditions which make its applicability to real world phenomena extremely limited at best. Neither of these aspects of research is calculated to aid one in deciding what the practical importance of particular variables might be in real life situations. As the demand from those involved in applied behavioral research increases, I suspect that we shall see some changes in these trends. When challenged "to put up or shut up," I find it hard to believe that behavioral scientists will be able to shut up.

Achieving a greater agreement about the important variables to study would help in clarifying decision rules for selecting questions, but it certainly would not go the whole way. Even among those who agree in general about the importance of particular variables, there can be serious division regarding the conceptual status of specific variables. Take, for example, a variable such as job satisfaction. Is this a single-dimensional, two-dimensional, or n-dimensional variable? Is job dissatisfaction the opposite end of a single scale, with job satisfaction at the other end, or should job satisfaction be conceptualized as consisting of a single general factor with some number (usually unspecified) of specific factors? Since job satisfaction is probably the area in which the greatest amount of research has been done, much of which has been conducted with fairly direct applied interest, one cannot be very optimistic about the chances for early clarification.

If one looks at notions of work motivation, the picture is even murkier. Is a motive conceived as a generalized energizer or as a "push" toward particular goals or activities? Are motives related to kinds of activities or to the value of certain outcomes? How do motives differ from occupational values? The ways in which these problems of conceptualization are resolved will have significant implications on how the variables are viewed in relation to the particular program being evaluated, and therefore on the kinds of questions selected. As with the problem of agreement on the relevant variables, one can only hope that the greater need for conceptual clarity will force behavioral scientists to think through more thoroughly the nature of their concepts and to do the necessary work to resolve the difficulties.

Let me now turn to two other related difficulties in the current status of the development of survey research methodology. These involve the apparent lack of interest in systematic methodological research and the underdevelopment of measurement theory in survey work as compared with the development of sampling theory. It seems a safe generalization that in the field of survey research, methodological research has a relatively low priority. This is not to say that there is no methodological research being conducted, but rather that the research that is done tends to be fragmented, local, unpublished, and usually specific to particular studies. The bulk of methodological research appears to be done for the internal benefit of large organizations, such as the Census Bureau, the Survey Research Center at the University of Michigan, and NORC. I expect that other academically based survey organizations, many commercial and market research firms, and individual scholars scattered around the country have also done important methodological work centered on their particular concerns. However, for the most part this work does not get published. With very few exceptions, systematic methodological work is notable by its absence.

It is easy to see how this lack of interest is perpetuated. There is comparatively little professional payoff (at least in the behavioral sciences) for good methodological work in the area of variable measurement. Such payoff currently goes to those who work on techniques of data analysis and in sampling theory. It is more difficult to discover why the reward system should operate in this fashion, because one would think that the value of sophisticated analytic techniques would be negated by the poor quality of the data being analyzed. Behavioral scientists appear to have a curiously ambivalent attitude regarding their data. Sometimes they take the stance that they know perfectly well that most of the data they use are of very poor quality and have a high amount of error, but they believe that sophisticated analytic techniques will enable truth to shine through all the data noise. On the other hand, at times they write and publish as if they were blissfully unaware that any serious measurement problems existed in the data and as if they need to be concerned only with sampling error, and sometimes not even with that.

How does one explain this comparative lack of interest in response errors as compared with the fairly sophisticated development of psychometric models in the measurement of educational achievement and individual differences in abilities? I suspect, although I have no evidence, that the difference lies in the uses to which data in social research have been put, compared with the ways in which data are used in educational systems and in personnel selection and placement. For the most part, survey data in the behavioral sciences have not been used to make important decisions concerning people's lives. There is every indication, however, that survey data will play an increasingly important role for social-planning purposes and in the evaluation of social programs. In addition, the allocation of large sums of money is now and will increasingly in the future be influenced by the results of sample surveys. Unfortunately, I believe that we are presently at a state where the quality of most survey data is too poor to support the uses to which it is put. We must work to improve very significantly the quality of data that we are collecting.

There are a few signs that things are beginning to change for the better. One of the most promising is the publication by the Institute for Social Research at the University of Michigan of two volumes that attempt to survey the state of measurement in the areas of political attitudes and occupational attitudes and characteristics.[1] John Robinson and his colleagues bring together in one place, to my knowledge for the first time, many different measures of the same variable, and attempt to review systematically the adequacy and usefulness of each measure. In many respects these volumes parallel those that have existed for years in the area of ability and achievement tests. It is perhaps an eloquent comment on the state of concern for methodology in survey work that such volumes should be only just now forthcoming. I trust that this is the beginning of a series which will be continuously improved and updated.

The authors of these volumes have used three groups of criteria to evaluate the scales reported. These are: (1) criteria relating to item construction, such as the sampling of relevant content, item analyses, and adequacy of question wording; (2) response set criteria, that is, assessing the attention paid to problems of response bias, acquiescence set, etc.; and (3) psychometric criteria, such as measures of reliability, normative information on the scale, and the ability of the scale to discriminate between groups known to differ in the dimension of concern. The fact that these sets of criteria have been applied consistently to the evaluation of the scales presented in these volumes adds immeasurably to their usefulness. The overwhelming impression that one receives from reading these volumes is that even in areas which have been fairly well studied, such as the measures of job satisfaction, the level of scale development work is surprisingly low and the information necessary for adequate evaluation of many measures is simply not available. I hope that the publication of these studies in systematic form will help rectify this situation.

Finally, I would like to mention a set of problems that are extremely disturbing to anyone seriously interested in the methodology of question asking--particularly disturbing because they seem so intractable. These difficulties stem from changes in linguistic usage across time and variations in usage among different subgroups within the same population or across different populations. In short, these are the vexing problems of the comparability of question meaning to the respondents at different points in time or at different points in space. Shifts in the use of language over time is a particular problem if one is interested in monitoring social change, such as changes in attitudes toward certain types of programs (for example, social security or social welfare programs), or is concerned with measuring changes in racial attitudes. To give just one example of the kinds of shifts that occur in word usage, in a study of racial attitudes in 1950 NORC asked white residents of neighborhoods in Chicago the following question: "Do you approve or disapprove of white and colored children being in the same schools together?" In 1964 we were asking the question in this form: "Do you think white students and Negro students should go to the same schools or to separate schools?" I expect that in the near future we shall be asking the question in terms of "Do you think white students and black students should go to the same or separate schools?" These alterations in wording are small and may be inconsequential adjustments to shifting usage. But on the other hand, we don't really know what are the effects of the changes. We all know that slight variations in the wording of a question may bring about relatively large shifts in the distribution of responses, but we have no systematic data that would allow us to approximate how much of the change in distributions is due to alterations in question wording and how much is due to real changes in opinion.

It is perhaps fortunate that we do not have much that can pass for trend data in social or political attitudes, since the problem of change in question wording over time has not really been seriously faced. However, if we should begin to rely on surveys for significant social indicator data, much more serious consideration will have to be given to this problem.

The problem of wording differences is not confined to changes over time but is also omnipresent in cross-sectional research. We have taken it more or less as a canon of faith in survey research that all respondents--all those who speak English at least--should be asked the questions in exactly the same wording, regardless of their educational level. The result of this article of faith is that questions addressed to nationwide samples are couched in a vocabulary that is presumed to be understandable by even poorly educated respondents. Such a presumption, however, may have little basis in fact for we know little about the way in which poorly educated or minority-group respondents interpret questions worded in standard, albeit simplified, English. Many of us in survey research feel that there is probably considerable loss of information when the identical question wording is used for both middle-class and extremely poor or minority-group respondents. But we have not found any usable way to alter question wording so that it becomes appropriate to the characteristics of the respondent. On the other hand, I don't think we have tried very hard either.

Recently the Social Science Research Council's Committee on Sociolinguistics sponsored a conference on "Language as Obstacle and as Data in Sociological Research." Participants in this conference pointed out that we need to be concerned not only with the problem of the meaning of words and sentences in interview schedules, but we should also pay some attention to the context within which interrogative sentences are embedded. The SSRC group called for more attention to what they termed "the ethnography of asking questions." In his report on this conference Allen Grimshaw noted:

> We simply do not know how to phrase questions that will be meaningful to random samples of diversified populations. We suspect that fixed-choice questions should never be used in comparative studies. Since those who use them have not systematically examined the possibly resulting biases, however, we have no way of estimating the magnitude or direction of errors

that are thereby introduced. . . .
Social scientists must ask themselves some serious questions about how wisely they ask questions of their research subjects. Otherwise, continuing refinements in quantitative analysis of data will produce only spurious or at best marginal increments of socially and sociologically relevant data.[2]

This is a view that I believe deserves very serious consideration by all those who are active in the area of survey research.

From the foregoing discussion of five obstacles to the delineation of precise decision rules for selecting questions, we can now point to some rules of thumb for use in asking questions about what questions to select. I would summarize these rules of thumb as follows:

1. What is the theoretical relevance to the problem at hand of the social-psychological variables that I select for study? If I cannot specify what these variables are and at least what their theoretical relation is to the phenomena I am studying, then I should abandon the effort to measure them.

2. From the best information I can get on operational measures of these variables, what measure best meets the criteria of good item construction and least susceptibility to response biases and has the best psychometric properties? It may turn out at this point that there are no scales which meet the minimum standards I have set for my research, and I must either abandon the effort to measure the variables or embark on a side excursion in developing new measures.

3. Finally, what are the characteristics of the population that I am surveying, particularly with regard to degrees of heterogeneity which might require different forms of questions for different segments of the population to be studied? While I believe it unlikely that one will find any measure that will have alternate forms of questions for differing subgroups of the population, I believe that the researcher should seriously ask himself whether he should not devote some of his precious research time to investigating the potential biasing effects of using the same question form for all respondents.

I have no illusions that the practice of survey research comes anywhere near to approximating the ideal toward which we strive. I do, however, feel strongly that we must make very substantial improvements in our measurement standards if we are to fulfill the promise that survey research methodology has made to those who are engaged in social research.

FOOTNOTES

[1]John P. Robinson, Jerrold G. Rusk, and Kendra B. Head, Measures of Political Attitudes (Ann Arbor: Survey Research Center, Institute for Social Research, University of Michigan, September, 1968); John P. Robinson, Robert Athanasiou, and Kendra B. Head, Measures of Occupational Attitudes and Occupational Characteristics (Appendix A to Measures of Political Attitudes) (Ann Arbor: Survey Research Center, Institute for Social Research, University of Michigan, February, 1969).

[2]Allen D. Grimshaw, "Language as Obstacle and as Data in Sociological Research," Items, v. 23, no. 2 (June, 1969), p. 21.

HYPOTHETICAL QUESTIONS AS MEASURES OF LABOR MOBILITY*

Herbert S. Parnes and Ruth S. Spitz, The Ohio State University

Introduction

In a dynamic economy, optimal allocation of resources depends upon the mobility of the factors of production. With respect to labor, this means that workers must be willing to move among jobs in response to perceived differentials in "net economic advantage." For example, if the economy were purely competitive, wage differentials for a given type of work would signify differences in contribution to social product. With perfect mobility, workers would respond to such differentials by moving from where their productivity is low (low-wage firms) to where it is higher (high-wage firms), and this flow would cause the wage to rise in the firms that are losing labor and to fall in those which are adding labor. When wage equalization occurred, labor of the given type would be optimally distributed among firms, and there would be no further stimulus to movement. Thus, in the theory of a competitive labor market, the processes of wage determination and labor allocation are intimately linked. 1/

The crucial role of labor mobility in imparting this kind of flexibility to the economy has stimulated a considerable amount of research on the subject, an important objective of which has been to ascertain the extent to which actual labor markets resemble the competitive model. For example, students of the labor market have attempted to ascertain how mobile the labor force is, who the mobile and immobile workers are, how much knowledge about alternative jobs workers have, the factors they take into account in making job decisions, and whether movement is actually in the direction of higher-wage jobs. 2/ In virtually all of this research, mobility has been measured by the frequency of job changes or by the length of service in a given job assignment.

While data on actual job changes are important for some purposes, they leave something to be desired as measures of mobility in the context of labor market theory. In that context mobility refers to the propensity of workers to make job changes in response to a perceived "net economic advantage" in doing so. The actual moves that workers make may or may not be reflections of such propensities. Job changes occasioned by layoff or discharge are clearly irrelevant to mobility in the above sense. Even if attention is confined to voluntary job changes, differentials in movement may reflect differences in opportunities rather than in propensities to move. In this paper, we discuss a method of measuring mobility as a propensity to change jobs in response to economic incentives and present the results of such measurement for national samples of employed men in the age groups 16-24 and 45-59.

Source of data

As the first stage of a longitudinal analysis of labor market experience and behavior, interviews were conducted with national probability samples of about 5,000 men 45 to 59 years of age and of the same number of male youth 14 to 24 years of age in the summer and autumn, respectively, of 1966. 3/ In addition to the kinds of attitudinal variables reported in this paper, detailed information was collected on current labor force and employment status, previous employment experience, characteristics of current job or most recent job for those with work experience, education and training, health, assets, income, and labor market status of other family members. Also, questions on retirement expectations were asked of the older group of men and questions on educational plans and occupational aspirations were asked of the youth. Five annual follow-up surveys of the two samples are planned, two of which have already been conducted. 4/ The five-year record of the work experience of the two samples that will ultimately be available will permit the predictive value of our mobility measures to be tested.

A Conceptual Framework for Studying Mobility

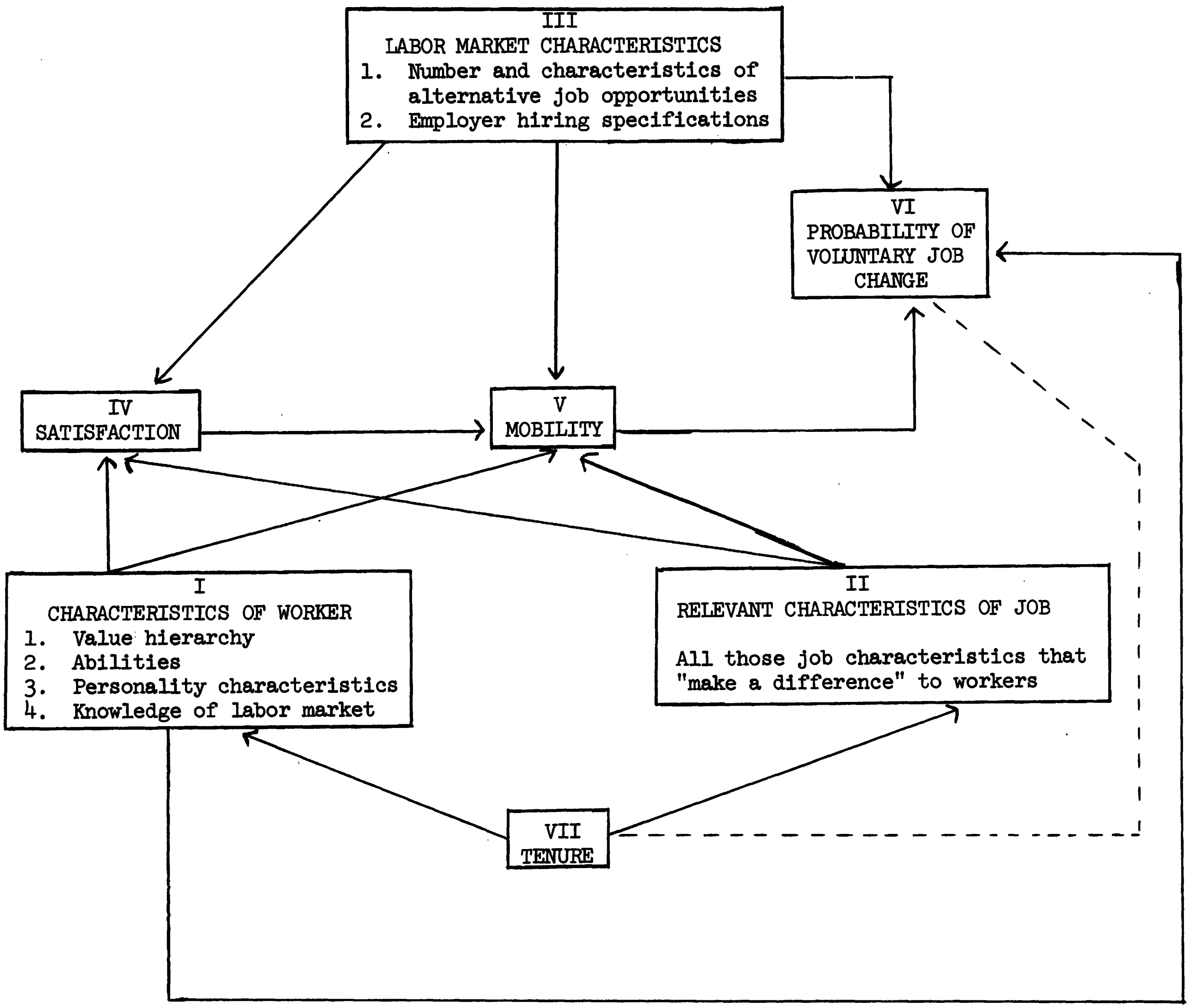
The relationships that we hypothesize among a worker's satisfaction with his job, 5/ his propensity to leave it for a more attractive alternative (mobility), and the probability of his actually making a voluntary job change may be represented schematically as shown in Figure 1, where the arrows are intended to indicate the

---

* The research reported here is part of an ongoing longitudinal study being conducted under contract with the Manpower Administration, U. S. Department of Labor under the authority of the Manpower Development and Training Act. A substantial portion of the paper is an adaptation of materials presented in the first report on that study: Herbert S. Parnes, et al., The Pre-Retirement Years (Columbus: The Ohio State University Center for Human Resource Research, October, 1968). Researchers undertaking such projects under Government sponsorship are encouraged to express their own judgments. Interpretations or viewpoints stated in this paper do not necessarily represent the official position or policy of the Department of Labor.

Figure 1

A CONCEPTUAL FRAMEWORK FOR STUDYING MOBILITY



direction of influence. A worker's satisfaction or dissatisfaction with his job is conceived to be produced by the interaction of three sets of variables: the characteristics of the worker (Box I), the characteristics of the job (Box II), and the characteristics of the labor market (Box III). For example, workers who differ in the relative values they attach to high income versus congenial personal relations would be expected, other things being equal, to register different degrees of satisfaction with a job in a small, low-wage firm. The characteristics of the labor market are also relevant, because they condition the worker's expectations. That is, the degree of satisfaction that a worker feels in a particular job depends in part upon the characteristics of other jobs he happens to know about. 6/

Mobility (Box V), like satisfaction, is conceived to be purely an attitude on the part of the worker. While it is related to satisfaction (note the arrow between IV and V), it is not exclusively a product of it, since the characteristics of the worker, the job, and the labor market may combine to produce a high propensity to change jobs even when satisfaction is high or a low propensity even in the face of dissatisfaction. For example, a worker who places a high premium on security may be unwilling to sacrifice his seniority in his present job despite dissatisfaction with it on other grounds, while an equally dissatisfied worker to whom security is not as important would have fewer reservations about leaving. Or a worker with low satisfaction may nevertheless be reluctant to leave his job because of a personality structure that makes him fearful of entering a new and unknown environment. 7/ Since mobility refers to a worker's willingness to take another job that is presumably available, it might appear that the character of the labor market is irrelevant. But this is not the case. Since there is no assurance that any proffered job will be permanent, a worker's reaction to exchanging his present job for another will be influenced by his appraisal of the availability of other opportunities should he experience a layoff or should the new job for some other reason not meet his expectations.

The probability that a worker will actually make a voluntary job change (Box VI) depends not only on his propensity to respond to more rewarding opportunities (mobility), but also on the opportunities for movement provided by the labor market (Box III) and those personal characteristics that determine (a) the extent of his knowledge of alternative jobs, (b) his initiative and vigor in pursuing them, and (c) his attractiveness to other employers (Box I). Thus, high mobility does not result in movement unless there are more attractive jobs that the individual knows about and unless he is acceptable to other employers. Labor market characteristics, it should be noted, play a dual role in affecting the probability of a job change: they condition the worker's attitude toward his present job and possible alternatives (satisfaction and mobility) and, at the same time, determine the possibilities for actual movement.

The broken line between Boxes VI and VII is intended to signify that tenure is the retrospective or historical reflection of the probability of movement. The arrows between Box VII and Boxes I and II indicate that length of service in a job operates indirectly to affect satisfaction and mobility through its direct effect on the characteristics of the job and the worker, particularly the former. A job in which a worker has served for ten years differs from the same job when he first entered it both because his tenure creates important equities (e.g., protection against layoff, vacation and pension rights) and because psychological and sociological bonds tend to be stronger. The characteristics of the worker also are affected by tenure. For instance, his relative attractiveness to his employer as compared with his attractiveness to other employers may be expected to increase as the result of the development of skills that are more or less specific to the particular firm.

Measures of Satisfaction and Mobility

Satisfaction It is difficult to define job satisfaction operationally, 8/ unless one uses the approach suggested by Hoppock in his early work on the subject: job satisfaction is "any combination of psychological, physiological, and environmental circumstances that causes a person truthfully to say 'I am satisfied with my job.'" 9/ Our measure of satisfaction is just this simple. Respondents were asked "How do you feel about the job you have now? Do you like it very much, like it fairly well, dislike it somewhat, or dislike it very much?" From the standpoint of our present interest, this measure leaves something to be desired because it is not clear to what extent a respondent is reacting to the intrinsic nature of the work he does and to what extent his response reflects his attitude toward such extrinsic factors as rate of pay, hours of work, nature of supervision, etc. Since there are questions in the interview schedule on the factors liked best and factors liked least about the job, we shall eventually be able to do some experimentation to see whether differentiating between the two sources of satisfaction and dissatisfaction makes a difference so far as interfirm mobility is concerned.

As in virtually all previous surveys that have used a similar question, the vast majority of employed men -- about nine-tenths of the younger sample and a slightly larger proportion of the older group -- express positive attitudes toward their jobs, and negligible numbers -- only about 2 percent -- express strongly unfavorable attitudes (Table 1). 10/ If those who report only moderately favorable attitudes are grouped with the two categories expressing some degree of dislike of their jobs, we have two nearly equal groups -- those who like their jobs very much and all others.

Mobility In view of the definition of mobility as a propensity to respond to perceived economic differentials, we wished to devise a simple measure that would abstract from the respondent's knowledge of the job market and from the number of opportunities that might exist for him, given his characteristics and those of the local labor market area. Moreover, we wanted our measure to relate exclusively to propensity to change employers -- either within or outside the local labor market area. We therefore asked each respondent employed as a wage or salary worker the following question: "Suppose someone in this area offered you a job in the same line of work you're in now. How much would the new job have to pay for you to be willing to take it?" An identical question was asked relating to a hypothetical job somewhere outside the local area. In both cases, responses were coded in relation to current wage rates. Thus, employed males are classified in terms of the percentage increase in rate of pay that they report would be necessary to induce them to make (1) an interfirm shift in the same labor market area, and (2) a geographic shift to some other area of the country. The distributions for both the younger and older men are shown in Table 2.

As anticipated, there are substantial differences in this measure of mobility between the two age groups and, within each group, depending upon whether the reference is to a job

Table 1

Satisfaction with Current Job: Employed Men 16-24 [(a)] and 45-59 Years of Age, 1966
(Percentage distribution)

Degree of satisfaction	16-24 years of age	45-59 years of age
Like it very much	49	57
Like it fairly well	40	35
Dislike it somewhat	7	5
Dislike it very much	2	2
Not ascertained	1	1
Total percent	100	100
Total number (thousands)	5,828	13,895

NOTE: In this and all subsequent tables, absolute figures are population estimates based on the number of sample cases. Percentages may not add to 100 because of rounding.

(a) Excludes those enrolled in school.

Source: National Longitudinal Survey

Table 2

Reaction to Hypothetical Job Offer In and Outside of Local Area: Employed Male Wage and Salary Workers 16-24 [(a)] and 45-59 Years of Age, 1966

(Percentage distribution)

Reaction to hypothetical job offer	Job offer in local area		Job offer outside local area	
	16-24 years of age	45-59 years of age	16-24 years of age	45-59 years of age
Would accept at same or lower wage	20	13	9	8
Would accept for wage increase of less than 10 percent	8	4	3	1
Would accept for wage increase of 10-50 percent	40	28	24	14
Would accept for wage increase of more than 50 percent	9	9	22	14
Would not accept at any conceivable wage	15	35	29	50
Not ascertained	8	11	13	13
Total percent	100	100	100	100
Total number (thousands)	5,566	11,011	5,566	11,011

NOTE: See note, Table 1

(a) Excludes those enrolled in school.

Source: National Longitudinal Survey

change in the locality or to one that would require a change of residence. 11/ A fifth of the young men but only an eighth of the older group report a willingness to accept a job at the same -- or even a lower -- wage rate. At the other extreme, 35 percent of the older men but only 15 percent of the youth indicate complete immobility -- an unwillingness to take another job in the area at any wage rate. In reacting to a job outside the local area, the proportions of completely immobile workers become one-half and about three-tenths for the older and younger age groups, respectively.

We are aware of the limitations of hypothetical questions as predictors of actual behavior. 12/ For one thing, irrespective of the subject matter, it is virtually impossible to specify all of the hypothetical conditions that might affect an actual decision. In the present case, for example, what exactly is meant by the qualification that the hypothetical job is "in the same line of work?" How large is the firm and how stable are its operations? How conveniently is it located to the respondent's residence? These and many other factors would doubtless be considered by an individual in a "real life" situation. Moreover, it has also been observed that responses to hypothetical questions may simply reflect the respondent's view of how he ought to behave rather than indicate how he would behave. This objection seems less relevant to the present application, since it is doubtful that there are strong moral or ethical considerations involved in most labor market decisions of this kind.

In any case, it is not our intention to interpret any of these responses literally. It is not necessary to debate whether the substantial minority of older men who say they would not change jobs for any conceivable wage increase would really turn down a job paying a million dollars a year. Our only purpose is to categorize individuals according to their relative mobility, i.e., their propensity to respond to economic incentives in the labor market. We therefore assume only that workers who say that they would change jobs for a small (or no) wage increase are more mobile than those who would require a larger increase, and that the least mobile of all are those who assert that they can conceive of no wage increase that would induce them to change. In the analysis that follows, which is confined to a consideration of mobility within the local labor market, we designate as "mobile" those workers who report a willingness to take another job at any specified wage and as "immobile" those who report that they would not take another job at any conceivable wage.

Some Correlates of Mobility

Tenure and type of occupation The hypothesized relationship between length of service and mobility as measured by the hypothetical job offer question is strong in the case of the older group of workers (Table 3). While 64 percent of these workers with less than 10 years of service are mobile, only 43 percent of those with 20 or more years are. In the case of the younger group, there is virtually no such relationship (Table 4), probably because of the very limited variability in tenure among this group whose oldest members are only 24 years of age.

Among the younger group of workers there is no perceptible difference in mobility between white collar and blue collar workers -- about three-fourths of each are mobile. In the case of the older men, however, there is an interesting interaction between type of occupation and length of service. Among those with fewer than 20 years of service, blue collar workers manifest a consistently greater willingness to change employers for more money than do white collar workers. The interoccupation difference in the proportion of mobile workers is 7 percentage points both among those with less than 10 years and those with 10-19 years of service. However, the difference is reversed for those with over 20 years of service, among whom 45 percent of the white collar and 41 percent of the blue collar workers are mobile. Stating this another way, long tenure appears to have a more substantial influence in reducing the mobility of blue collar than of white collar workers, doubtless because of the greater relative importance of seniority in providing job security for the blue collar group.

Degree of job satisfaction Among the 45-59 year age group, degree of job satisfaction and mobility are related in the expected manner: men who say they like their jobs very much are less likely to be mobile than those who express lesser degrees of satisfaction (Table 5). The difference virtually disappears, however, for men with 20 or more years of service. In other words, degree of satisfaction and length of service exert independent influence on mobility only through part of the total range of length of service. Once a worker has accumulated enough service, the degree of satisfaction he feels in his job is almost irrelevant to his willingness to make a change. It is worth noting that the longest service workers who express some reservations about their jobs are nevertheless less mobile even than short service workers who express the highest degree of satisfaction.

The inverse relationship between degree of job satisfaction and mobility is also evident in the case of youth (Table 6). Among both white collar and blue collar employees between the ages of 16 and 24, less than three-fourths of those who are highly satisfied are classified as mobile, as compared with four-fifths of those who express lesser degrees of satisfaction with their jobs.

Table 3

Mobility,[a] by Length of Service in Current Job and Type of Occupation: Employed Male Wage and Salary Workers 45-59 Years of Age, 1966

Length of service and type of occupation	Total number (thousands)	Percentage distribution			
		Mobile	Immobile	Not ascertained	Total
Less than 10 years:					
White collar	1,536	59	27	14	100
Blue collar	2,479	66	23	10	100
Total [b]	4,630	64	25	11	100
10-19 years:					
White collar	1,016	48	38	14	100
Blue collar	1,664	55	38	8	100
Total [b]	2,916	52	38	10	100
20 or more years:					
White collar	1,360	45	43	12	100
Blue collar	1,850	41	49	10	100
Total [b]	3,397	43	46	11	100
Total employed: [c]					
White collar	3,932	52	35	13	100
Blue collar	6,030	55	35	9	100
Total [b]	11,011	54	35	11	100

NOTE: See note, Table 1.

(a) Based on response to hypothetical job offer. Respondents who specified a wage at which they would accept proffered job are classified as mobile. Those who said they would not accept the job at any conceivable wage are classified as immobile.

(b) Totals include 746 thousand service workers, 281 thousand farm workers, and 23 thousand individuals for whom occupation was not ascertained.

(c) Includes 68 thousand for whom length of service was not ascertained.

Source: National Longitudinal Survey

Table 4

Mobility,[a] by Length of Service and Type of Occupation: Employed Male Wage and and Salary Workers 16-24 Years of Age, Not Enrolled in School, 1966

Length of service and type of occupation	Total number (thousands)	Percentage distribution			
		Mobile	Immobile	Not ascertained	Total
Less than 1 year:					
White collar	616	75	12	13	100
Blue collar	1,972	78	15	7	100
Total [b]	2,937	77	14	9	100
1-2 years:					
White collar	528	76	15	9	100
Blue collar	1,049	78	16	7	100
Total [b]	1,724	77	16	7	100
3 or more years:					
White collar	215	77	17	6	100
Blue collar	576	76	15	8	100
Total [b]	895	76	16	8	100
Total employed: [c]					
White collar	1,359	76	14	10	100
Blue collar	3,598	78	15	7	100
Total [b]	5,566	77	15	8	100

NOTE: See note, Table 1.
(a) See footnote (a), Table 3.
(b) Totals include 355 thousand service workers, 199 thousand farm workers and 53 thousand individuals for whom occupation was not ascertained.
(c) Totals include 8 thousand workers for whom length of service was not ascertained.
Source: National Longitudinal Survey

Table 5

Mobility, [a] by Length of Service in Current Job and Degree of Job Satisfaction: Employed Male Wage and Salary Workers 45-59 Years of Age, 1966
(Percentage distribution)

Length of service and degree of satisfaction	Total number (thousands)	Mobile	Immobile	Not ascertained	Total
Less than 10 years:					
Like job very much	2,446	58	32	11	100
All other	2,143	71	18	11	100
Total [b]	4,630	64	25	11	100
10-19 years:					
Like job very much	1,670	47	42	11	100
All other	1,222	60	32	8	100
Total [b]	2,916	52	38	10	100
20 or more years:					
Like job very much	1,990	42	48	10	100
All other	1,377	44	44	11	100
Total [b]	3,397	43	46	11	100
Total employed: [c]					
Like job very much	6,141	50	40	11	100
All other	4,776	60	29	10	100
Total [b]	11,011	54	35	11	100

NOTE: See note, Table 1.

(a) See footnote (a), Table 3.
(b) Totals include 94 thousand workers for whom degree of job satisfaction was not ascertained.
(c) Total includes 68 thousand workers for whom length of service was not ascertained.

Source: National Longitudinal Survey

Table 6

Mobility, [a] by Type of Occupation and Degree of Job Satisfaction: Employed Male Wage and Salary Workers 16-24 Years of Age, Not Enrolled in School, 1966
(Percentage distribution)

Type of occupation and degree of satisfaction	Total number (thousands)	Mobile	Immobile	Not ascertained	Total
White collar:					
Like job very much	755	72	17	11	100
All other	605	81	11	9	100
Total	1,359	76	14	10	100
Blue collar:					
Like job very much	1,639	74	20	6	100
All other	1,960	81	11	8	100
Total	3,598	78	15	7	100
Total employed: [b]					
Like job very much	2,666	73	19	8	100
All other	2,898	81	11	8	100
Total	5,566	77	15	8	100

NOTE: See note, Table 1.

(a) See footnote (a), Table 3.
(b) Totals include 355 thousand service workers, 199 thousand farm workers and 53 thousand workers for whom occupation was not ascertained.

Source: National Longitudinal Survey

Relation Between Mobility Measure and Actual Job Movement

The pronounced relationships that have been found to exist between our measure of mobility and age, tenure, and degree of job satisfaction are precisely what one would expect of a variable that did, indeed, represent the propensity of a worker to change jobs. Nevertheless, the real test of the validity of the measure lies in its ability to discriminate, in the context of our conceptual framework, between those who do and those who do not make voluntary job changes. More specifically among the hypotheses that we shall wish to test as the work experience of our samples unfolds are the following: (1) workers whom we have classified as mobile are more likely than those designated as immobile to make voluntary job changes over the five years of the study; (2) among mobile workers, those in "tight" labor markets (low unemployment) are more likely to change jobs voluntarily than those in "loose" labor market areas; (3) controlling for occupation, the probability that a mobile worker will make a voluntary job change is positively related to (a) his education and training and (b) the extent of his knowledge about the labor market.

The only relevant tabulations of follow-up data that are yet available to us relate the mobility characteristics of the older group of men to whether they had changed employers between the 1966 and 1967 surveys. Table 7 shows the proportions of wage and salary workers in 1966 who made voluntary job changes between the two survey dates, classified according to length of service in their 1966 jobs and according to the degree of mobility they evidenced in the 1966 survey. 13/ Men who had been classified as mobile on the basis of their responses in 1966 to the hypothetical job offer question were more than twice as likely as those classified as immobile to have made voluntary job changes between the 1966 and 1967 surveys. This overall difference of about 4 percentage points is statistically significant.

Much of the relationship between our measure of mobility and the probability of actual movement is a reflection of the inverse relationship already described between the mobility measure and length of service in 1966 job, since it is well known that the probability of voluntary job separation is inversely related to length of service. It is important to inquire, therefore, whether the mobility measure discriminates between voluntary job changers and non-changers within length of service categories. The answer to this question appears to be affirmative. Although the differences in the voluntary separation rates of mobile and immobile workers are not large enough to be statistically significant at the 5 percent level, their consistency allows some confidence that they are real rather than simply reflecting sampling variation. As the number of persons who leave their 1966 jobs increases during the 5 years of the study, we shall probably be able to arrive at a more positive conclusion on this matter.

If one accepts the figures in Table 7 at their face value, the mobility measure shows a stronger relationship to voluntary job changing among long-service than among short-service workers. Among those with less than 10 years of service, men classified as mobile are only slightly more likely than those classified as immobile to have changed jobs (11.6 versus 8.6 percent); among men with 10-19 years of service, the mobile are three times as likely as the immobile to have changed jobs (3.4 percent versus 1.1 percent) and among those with 20 or more years of service, eight times as likely (2.4 percent versus 0.3 percent).

Stated another way, length of service appears to have an effect on the probability of a voluntary separation that is independent of the worker's propensity to move as measured by the hypothetical job offer question. This may mean that there are dimensions of mobility that the question does not measure. Alternatively, or additionally, it may reflect the fact that tenure is associated with characteristics of the worker that interact with the characteristics of the labor market in such a way as to cause opportunities for movement to be different for workers with different periods of service in their current jobs. The positive association between length of service and age is an obvious example. It is also clear that long service in a job makes a worker more valuable to the current employer, who is therefore more likely than he would be in the case of a shorter service employee to match an offer from a competing employer, and thus prevent the job change.

Conclusion

From a methodological point of view, our findings to date suggest that a question posing a hypothetical job offer can be used as a measure of the mobility of workers, defined as their propensity to change employers in response to a perceived economic advantage in doing so. Additional analysis of the data over a longer period of time and in relation to additional variables will be necessary before we can be confident that the measure provides a substantially better basis for predicting actual job changes than length of service alone would provide.

From a substantive point of view, perhaps the most important conclusion to be drawn from our findings thus far is that labor mobility is a much more complex phenomenon than would be imagined on the basis of conventional labor market theory, which tends to conceive of labor as a more or less homogeneous and fluid factor continuously flowing -- or at least oozing -- in the direction of net economic advantage. While this conception is doubtless valid and adequate for many purposes, it neglects the rich variety of behavior that actually exists in the labor market. A full understanding of how the labor market operates requires knowledge of who the mobile and immobile

Table 7

Number of Wage and Salary Workers Employed in 1966 and Proportions Making Voluntary Job Changes [(a)] between 1966 and 1967, by Length of Service and Mobility [(b)] in 1966 Job: Men 45-59 Years of Age in 1966

Length of service in 1966 job	Mobile workers, 1966		Immobile workers, 1966		Total [(d)]	
	Number (thousands)	Percent voluntary changers	Number (thousands)	Percent voluntary changers	Number (thousands)	Percent voluntary changers
Under 10 years	2,953	11.6	1,154	8.6	4,630	10.6
10-19 years	1,520	3.4	1,105	1.1	2,916	2.2
20 or more years	1,449	2.4	1,573	0.3	3,397	1.3
Total [(c)]	5,967	7.1	3,848	3.0	11,011	5.4

NOTE: See note, Table 1.

(a) The data probably understate slightly the number of job changers for two reasons. First, 5.8 percent of the 1966 sample were not re-interviewed in 1967, 1.7 percent because they could not be located. The latter group probably includes a disproportionately large number of job changers. Second, as the result of a tabulation error, persons who shifted from a wage or salary job in 1966 to self-employment in 1967 are not included among the job changers.
(b) See footnote (a), Table 3.
(c) Totals include 68 thousand workers for whom length of service was not ascertained.
(d) Totals include 1,196 thousand workers for whom mobility was not ascertained.

Source: National Longitudinal Survey

workers are and of the various characteristics of individuals and of the environment that condition their responses. The framework for analysis presented in this paper offers some hope of shedding additional light on questions of these kinds.

FOOTNOTES

1 For a fuller treatment of conventional labor market theory and a critical evaluation of it, see Lloyd G. Reynolds, The Structure of Labor Markets (New York: Harper and Brothers, 1951).

2 For a review of the literature, see Herbert S. Parnes, "Labor Force: Markets and Mobility," International Encyclopedia of the Social Sciences (New York: The Macmillan Company and The Free Press, 1968), Vol. 8, pp. 481-487.

3 The samples were designed by the U.S. Bureau of the Census, and interviewing was done by the same Census enumerators who are responsible for the Current Population Survey. For a description of the sampling, interviewing, and estimating procedures, see Herbert S. Parnes, Belton M. Fleisher, Robert C. Miljus, Ruth S. Spitz and Associates, The Pre-Retirement Years: A Longitudinal Study of the Labor Market Experience of the Cohort of Men 45-59 Years of Age (Columbus: The Ohio State University Center for Human Resource Research, October, 1968) Volume 1, Appendix B, and Herbert S. Parnes, Robert C. Miljus, Ruth S. Spitz and Associates, Career Thresholds: A Longitudinal Study of the Educational and Labor Market Experience of Male Youth 14-24 Years of Age (Columbus: The Ohio State University Center for Human Resource Research, February, 1969) Volume 1, Appendix B.

4 Comparable studies are being made of two age groups of women: 14 to 24 and 30 to 44 years old. No tabulations for either of these are as yet available.

5 We use "job" in this paper to refer to an affiliation with a particular employer rather than to service in a particular occupation, although we believe that the conceptual framework is equally relevant to occupational mobility.

6 One study has found that among workers with equal starting salaries, those whose previous jobs had paid more had higher quit rates than those who had come from lower-paying jobs. Reported in Frederick Herzberg et. al., Job Attitudes: Review of Research and Opinion (Pittsburgh: Psychological Service of Pittsburgh, 1957), p. 106.

7 See Nancy Morse, Satisfactions in the White Collar Job (Ann Arbor: University of Michigan Press, 1953), p. 53.

8 For a recent review of the literature on job satisfaction which stresses the difficulties of definition, see Bonnie Carroll, Job Satisfaction, Key Issues Series - No. 3, (Ithaca: New York State School of Labor and Industrial Relations, February, 1969), pp. 2-3.

9 Robert Hoppock, Job Satisfaction (New York: Harper and Brothers, 1935), p. 47.

10 It seems doubtful that many workers can psychologically afford to admit (even to themselves) dissatisfaction with a situation in which they apparently "choose" to remain. "Well, I guess I'm satisfied or else I wouldn't have stayed here, would I?" is not an atypical response by manual workers to a question on job satisfaction.

11 Because the samples are not purely random, conventional tests of statistical significance cannot be used. Using a formula suggested by Bureau of the Census statisticians, we have constructed charts which indicate roughly, for different ranges of bases and different magnitudes of the percentages themselves, whether a measured difference between two percentages may be considered to be significant at the 5 percent level. (See The Pre-Retirement Years, op. cit., Appendix C, pp. 258-266; Career Thresholds, op. cit., Appendix C, pp. 213-221). In this paper, we attach appropriate qualifications to any findings that do not meet this test of significance.

12 See, for example, Claire Selltiz, Marie Jahoda, Morton Deutsch, and Stuart W. Cook, Research Methods in Social Relations (New York: Holt and Company, 1960), p. 250; John Dollar, "Under What Conditions Do Opinions Predict Behavior?", Public Opinion Quarterly, Winter, 1948, pp. 628-632; S. L. Payne, The Art of Asking Questions (Princeton: Princeton University Press, 1951), p. 236.

13 These data probably slightly understate the proportions of wage and salary workers in 1966 who made job changes. For one thing, 5.8 percent of the 1966 sample were not re-interviewed in 1967, including 1.7 percent who could not be located, among whom there were probably a disproportionate number of job changers. Secondly, men who shifted from wage and salary jobs to self-employment have inadvertently been excluded from the job changers.

X

COMPUTER ASSISTED SIMULATION IN THE SOCIAL SCIENCES

Chairman, JAMES BESHERS, Queens College

EFFECTS OF ALTERNATIVE POLICIES FOR ALLOCATING FEDERAL AID FOR EDUCATION TO THE STATES*

Marilyn E. Manser, Thomas H. Naylor, and Kenneth L. Wertz, Duke University

Introduction

Two distinct lines of inquiry have been pursued in the analysis of state and local governmental expenditures. In his recent work, Governmental Problem Solving: A Computer Simulation of Municipal Budgeting, Crecine represents the methodological split as follows:

> The assumption that participants in the budget-making process are passive instruments who will come up with a predetermined solution to the problem of (public) resource allocation...by following economic dictates and service demands... vs. the assumption that budget-makers are organizational decision-makers and problem-solvers who structure complex problems, generate alternatives, and make choices on the basis of some criteria, is a real difference (3, p.20).

The first approach has informed a vast literature which searches among socioeconomic factors such as income and population density for determinants of public expenditures. The second and newer approach has stimulated investigations of the decision-making behavior of public officials.

It is the relevance of much of this latter evidence to states' expenditures for education which the present paper explores. After developing and estimating a model suggesting the manner in which such outlays might be set, we consider via simulation some of the model's normative applications to the problem of allocating federal aid for education among the states.

Development of the Model

As mentioned above, the focal point of this study is the behavior of participants in the budgetary process. It will be offered herein that at least three factors are prominent to the state officials responsible for drafting and approving the budget for a state's expenditures on education: (a) the amount spent on education by the state in the previous year; (b) the amounts spent on education by neighboring states in the previous year; and (c) the amount of federal payments for education to the state in the previous year. All factors will be regarded on a per capita basis. These variables are discussed in turn and in reference to the model (1):

$$(1)\quad S_t^i = a^i S_{t-1}^i + b^i N_{t-1}^i + c^i G_{t-1}^i + E_t^i$$

The superscript i refers to one of the forty-eight coterminous states; the subscript t marks annual steps, t = 1951,...,1965.

(a) Past expenditures by the state, S_{t-1}^i

An entirely rational approach to the allocation of public funds among public uses can be achieved in theory when one possesses a thoroughly defined system of values, certain knowledge of the consequences of a completely defined set of alternatives, and certain knowledge of their attendant costs. In practice, one attempting this comprehensive review of a budget finds "the absence of any single operational measure of efficiency in the public sector, uncertainty as to the result of alternative courses of action, and (one's limited capacity) to process the necessary information," (5). Davis, Dempster, and Wildavsky, following the lead of Wildavsky's work, observe that participants in budgeting, in reaction to the realities above,

> deal with their overwhelming burdens by adopting aids to calculation. By far the most important aid to calculation is the incremental method. Budgets are almost never actively reviewed as a whole in the sense of considering at once the value of all existing programs as compared to all possible alternatives. Instead, this year's budget is based on last year's budget, with special attention given to a narrow range of increases or decreases. Incremental calculations proceed from an existing base, (4, pp. 529-530).

This observation, registered again in Crecine's study (3) of municipal budgeting in three large cities and again in Gerwin's examination (5, 6) of budgeting in a municipal school system, promotes the appearance of past expenditures for education as an ingredient of the decision to be made with respect to current expenditures (7).

Sharkansky's use (14) of past expenditures as an explanatory variable provoked Harlow (8) to assert that while it probably enhanced the prediction of a given state's expenditures, it contributed little to the explanation of why states differ in per capita expenditures at a given point in time to transfer that question to an earlier year. It must be evident that over a long period of time, public officials do not have sole control over the magnitude of public expenditures; extraordinary events such as a prolonged depression or a total war visibly disrupt the stability of the process as it has been described here. Other economic and sociological shifts may also impinge upon public administrators. But the theoretical and empirical work cited above provides important material for the suggestion at least that short run (this study suggests about a decade by implication, though that matter is unclear) differences in per capita expenditures reflect in part short run differences in administration. Or again, it is entirely possible that a governmental budgetary system can operate according to its own devised rules within rather loose constraints presented by external forces until the products of this procedure induce a revision of the constraints (3). Rather than forcing an exclusive choice of the two approaches, the one focusing on socio-economic characteristics and the other on the budgetary process, or designating one as explanatory and the other as predictive, it would seem more advantageous to work towards their union. Perhaps that task can proceed when both modes of inquiry are more fully explicated.

(b) Expenditures in neighboring states, N^i_{t-1}

In lieu of measures of appropriateness (e.g., efficiency) and full examination of alternatives and their costs, another way in which state officials can monitor their actions is by surveying what their colleagues in other jurisdictions are doing. Benson (1) has stressed the influence of the process of comparison upon expenditures for education. Walker's study (16) of the adoption of innovations by state governments also finds the hypothesis of comparison as a decision-making aid to be useful and outlines some of the mechanisms by which competitive and emulative tendencies are spread among administrators from state to state.

Although pertaining to the local level, Gerwin's work provides a striking illustration (6). He discovered that a general increase in minimum salaries for teachers holding a bachelors degree in the Pittsburgh school system was granted only when Pittsburgh had dropped into last place among eighteen major cities in the Northeast (i.e., the systems with which Pittsburgh was most likely in competition for teachers). This decision rule was exercised on four different occasions between 1953 and 1964.

In model (1), the process of comparison is specified as follows:

$$N^i_t = \sum_{j=1}^{48} w^i_j S^i_t$$

where w^i_j equals the fraction of state i's land perimiter that adjoins state j. Portions of a state's boundary which touch the sea or a foreign country are not regarded as contributing to its land perimiter. Naturally if state i does not border on state j, then $w^i_j = 0$. The variable is lagged one period in recognition of the delay which occurs in the perception and availability of this information.

This specification of the comparison process is admittedly primitive, but shall suffice until additional research suggests a superior measure.

(c) Federal payments, G^i_{t-1}

Another occasion for official consideration of a departure from the base of the budget is the receipt of federal funds for education. Although these constitute a rather small portion of the total outlay S^i (about 9.6% in 1965), they do offer some flexibility, as many writers have noted. Because educational grants in aid are largely free of matching requirements (13, 15), administrators may choose to substitute federal funds for state funds; or simply spend the amount received; or expand their own effort by spending additional sums in excess of the aid received ("stimulation"). Since the S^i series are not net of the G^i series, one could identify the probable existence of one of the three alternatives above by examining statistically significant departures of c^i, the estimated coefficient of G^i, from one, although the regression analysis can only suggest, not prove, this interpretation.

The variable is lagged one period, again reflecting the fact that some time is required to accumulate and to evaluate the aid which had been received. In so far as current expectations about federal aid are not based on previous assistance, one would like to account for them; but no satisfactory formulation could be devised.

The specification of model (1), coupled with the nature of grants in aid to education, afford an escape from the brunt of the criticism (13) directed at the use of G^i. On the other hand, it does pose a degree of interdependence between S^i_{t-1} and G^i_{t-1}. Another potentially distortive econometric problem is to be recognized in the presence of the lagged variable S^i_{t-1} in that autocorrelated disturbances may appear. Consequently, a modified version of the Cochrane-Orcutt iterative technique (12) was employed for estimation of the parameters in order to meet this problem.

This section sought to describe an alternative, but not necessarily competitive methodological basis for analyzing state expenditures for education and to develop a model suggested by it and by the theoretical and empirical contributions cited above. The next section reports the outcome of the estimation of the parameters.

Estimation: Results and Discussion

The estimations of model (1) for each of the forty-eight coterminous states, using time-series data,[1] are reported in Table 1. Asterisks denote that the coefficients a^i and b^i are statistically different from zero and the coefficient c^i from one, at the .10 level. For the most part, the coefficients a^i and b^i fall in the zero-one interval, though this result was not at all anticipated for the latter. In two of the four instances in which a^i is negative, it is significantly so, thereby forcing one to reject the model's relevance for those two states. Thirty-eight of the remaining forty-six states find the a^i to be significantly different from zero, which evidence is consistent with the concept of a budgetary base.

Also, more than half of the b^i show significance. This finding is compatible with the operation of the comparison process mentioned earlier, though it does suggest that it is not as widespread as the text has urged. Comments on this head must be tempered, however, by recalling the tentative quality of the specification of N^i.

The estimated c^i cover a broad range. That some are less than zero indicates that the frequently-aired charge (particularly in reference to urban renewal and manpower retraining) of federal assistance tending to constrict the present expenditures of lower governments because the latter shall expect more generous sums in the

Table 1. Regression Results, Model of State Expenditures for Education, by State

State	Regression coefficients a, b, c: a	b	c	Standard Error	R^2
1. Ala.	0.311*	0.719*	1.003	3.074	.959
2. Ariz.	0.997*	0.071	-0.006	5.746	.938
3. Ark.	-0.225	1.120*	-0.418*	2.576	.953
4. Calif.	1.080*	-0.067	0.900	2.517	.977
5. Colo.	1.048*	0.036	0.189	3.180	.981
6. Conn.	0.895*	0.487	-4.977*	2.542	.956
7. Del.	0.304	3.257*	-7.311	16.589	.754
8. Fla.	0.628*	0.392*	0.742	3.822	.919
9. Ga.	0.296*	0.734*	1.474	2.352	.967
10. Idaho	0.717*	0.300*	-1.486*	2.232	.978
11. Ill.	-0.688	0.917*	6.699*	2.496	.967
12. Ind.	0.711*	0.399	1.672	2.635	.982
13. Iowa	0.734*	0.270	0.584	1.973	.979
14. Kan.	0.350	0.774*	-0.345*	2.970	.953
15. Ky.	1.049*	0.274*	-3.225*	3.678	.967
16. La.	1.184*	-0.168	-0.159	4.073	.953
17. Maine	0.954*	0.052	1.829	2.177	.968
18. Md.	0.846*	0.105	1.430	3.651	.948
19. Mass.	0.033	0.493*	2.474	1.766	.952
20. Mich.	0.694*	0.986*	-2.679*	3.310	.966
21. Minn.	0.739*	0.429	0.656	2.329	.987
22. Miss.	0.558*	0.547*	-1.067	4.337	.938
23. Mo.	-0.089	1.127*	-1.254*	2.266	.967
24. Mont.	0.306*	0.632*	1.725	3.951	.931
25. Neb.	0.723*	0.177	0.172	1.697	.939
26. Nev.	0.488*	1.016*	-6.784*	8.578	.899
27. N. H.	0.499*	0.438*	-0.019	2.078	.946
28. N. J.	0.833*	0.189*	-2.709*	2.147	.936
29. N. M.	0.895*	0.309	0.302	9.761	.897
30. N. Y.	1.135*	0.183	-4.646*	2.616	.986
31. N. C.	0.490*	0.555*	2.278	3.530	.951
32. N. D.	-0.577*	1.532*	3.479*	4.596	.947
33. Ohio	0.929*	0.086	-0.097	1.086	.978
34. Okla.	0.581*	0.566*	0.223	3.796	.938
35. Ore.	0.230	0.730*	0.336*	1.907	.992
36. Pa.	0.121	1.426*	-10.440*	3.683	.935
37. R. I.	0.809*	0.301	0.850	4.353	.926
38. S. C.	1.139*	-0.398*	6.430*	6.354	.508
39. S. D.	0.727*	0.181	1.112	2.121	.972
40. Tenn.	0.858*	0.272*	-1.208*	1.055	.992
41. Tex.	1.173*	-0.003	-1.810*	3.751	.932
42. Utah	0.417*	0.468*	2.567*	5.733	.969
43. Vt.	0.605*	0.674*	-0.090	5.243	.936
44. Va.	0.270*	0.663*	-0.268*	1.044	.991
45. Wash.	0.654*	0.792	-0.768*	3.930	.983
46. W. Va.	1.001*	0.050	0.513	2.402	.966
47. Wisc.	0.380	0.246*	3.548*	1.762	.991
48. Wyo.	0.623*	0.424	2.940	8.188	.828

future, may have some merit with respect to educational expenditures. Despite a number of large deviations from one, many of these are not statistically significant at the .10 level, which may be due to the larger sampling variances associated with the mild interdependence of S^i_{t-1} and G^i_{t-1}. Notwithstanding this disappointing feature of the c^i, the model (1) otherwise seems to perform well as judged by the range and significance of many of the estimated parameters and by the generally high R^2. The median value of the latter is .96.

The Policy Alternatives

To this point, our efforts in developing and estimating the model (1) have aimed at description. In the present section the emphasis shifts to a normative application of the model, suggesting the potential of using it, or a refined version of it, to guide federal policymakers in their disbursement of aid to education at the state government level.

Suppose that federal disbursements for education were to occur systematically according to one of the plans listed below. Though political considerations and legislative stipulations will make complete adherence to a single plan impossible, the adopted plan shall serve as a standard to be approximated more closely as the allocators' discretion is increased. Let us assume that the plan which is adopted is the one which seems to promise the greatest effectiveness in stimulating educational expenditures by the states as a group over the planning period in question. Naturally this assumption slights the vector of goals which in fact may be operative; but it also captures much of the intent of the federal grant program. The measure of effectiveness to be used is the ratio of total state educational expenditures to total federal payments to the states for education over a period of n years, the longevity of the plan:

$$Z = \sum_{t=1}^{n} \left(\sum_{i=1}^{48} S^i_t P^i_t \right) / \sum_{t=1}^{n} \left(\sum_{i=1}^{48} G^i_t P^i_t \right)$$

where P^i_t denotes the population of the i^{th} state in year t.

The six allocating schemes that shall be reviewed here are suggested more by the model than by observation of current procedures, although plans 1 and 2 are imitations of the "mark-up" rule of thumb which other observers (4) believe operative in the appropriation of funds to some federal agencies.

Plan 1. This "control plan" calls for annual increases in federal payments to state i at the rate g_i, the average growth rate of G^i for the past five years:

$$G^{i*}_t = (1 + g_i)\, G^i_{t-1}$$

Plan 2. Under this plan, the federal payment to each state is increased by 3% of the control plan payment:

$$G^i_t = (1.03)\, G^{i*}_t$$

Plan 3. Suppose that states that appear to have had their own expenditures for education stimulated by the receipt of federal funds were rewarded with an additional 7% of their control payments, while the remaining states were penalized by 2% of their control payments:

$G^i_t = (1.07)\, G^{i*}_t$ for c^i greater than 1

$G^i_t = (.98)\, G^{i*}_t$ for c^i otherwise

Plan 4. This plan is similar to plan 3, but it does not penalize states as long as their expenditures do not appear constricted by additional federal funds:

$G^i_t = (1.03)\, G^{i*}_t$ for c^i greater than 1

$G^i_t = (.98)\, G^{i*}_t$ for c^i less than 0

$G^i_t = G^{i*}_t$ for c^i otherwise

The narrow construction of these plans should make it unnecessary to remark upon the exploratory nature of this research. If more sophisticated variants of plans 3 and 4 were to be considered as policy components in a serious way, then one would want to regard the sampling variances of the c^i in addition to the c^i themselves.

Plan 5. Another interesting heuristic involves awarding additional grants to states whose neighbors seem influenced by the comparison process. To each of the twenty states having at least 60% of its border adjoining states showing significant coefficients b^j, give an extra 6% of the control payment; to the others, give the control payment only.

Plan 6. As a variation of plan 5, award an extra 3% of the control payment to each of the thirty-one states having at least 50% of its boundaries adjoining states showing significant coefficients b^j; to the others, give the control payment only.

Policy Simulations

To test the effects of the six allocation plans on Z, we conducted policy simulation experiments with the model for a seven-year period (1966-1972). For a given plan, one solves for S^i each period in terms of (a) the value of S^i generated by the model in the preceding period, (b) the value of the neighborhood variable, (c) the value of G^i_{t-1} as given by the particular plan, and (d) the stochastic error term. We assumed that the error terms were normally distributed with expected value equal to zero and standard deviation equal to the standard error of the estimate. For each plan we ran the simulation seven years and calculated Z. Population projections were generated by

$$P_t^i = (1+p_i)\, P_{t-1}^i$$

where p_i is the average growth rate of P^i for 1961 through 1965.

There are two reasons for making the simulations stochastic. First, by including the stochastic variable E_t^i, we take into consideration the random effects which have not been explained by the model. Second, we can now say something about the degree of confidence that we have in any inferences which we might make about the differences in the effects of the six policies on the state-federal educational expenditure ratio.

The simulation experiment was replicated thirty times per plan. The sample means, $\bar{Z}_j$, appear in Table 2.

Table 2. Sample Means, $\bar{Z}_j$

Plan j	$\bar{Z}_j$
1	7.1281
2	6.1762
3	7.4247
4	7.4272
5	6.2519
6	6.4423

Data Analysis

To begin analysis of the output data generated by the simulations it was asked whether the expected values of Z_j for the six allocation plans are equal; and if they are not, between what plans will one find differences? An F-test was conducted to test the null hypothesis that the state-federal educational expenditure ratios for the plans are identical. As the computed F of 1350 dwarfed the tabulated F with 5 and 174 degrees of freedom at the .005 level (3.35), the null hypothesis was discarded. In order to identify where the suggested differences might be, Tukey's method (10, 11) for constructing simultaneous confidence intervals was used; as the confidence allowance at this level is $\pm$ .0064, all the differences between the sample means $\bar{Z}_j - \bar{Z}_k$ (see Table 3) are significant except for those of plans 3 and 4.

Table 3. Differences of Sample Means, $\bar{Z}_j$-$\bar{Z}_k$

j \ k	2	3	4	5	6
1	0.9519	-0.2966	-0.2991	0.8762	0.6858
2	---	-1.2485	-1.2510	-0.0757	-0.2661
3	---	---	-0.0025	1.1728	0.9824
4	---	---	---	1.1753	0.9849
5	---	---	---	---	-0.1904

The sample means indicate that the federal government's contribution to total state spending over the simulation period was between 13.5% (plan 4) and 16.2% (plan 2) after beginning in 1965 at 9.6%. The simulations also suggest that the most effective way examined to induce additional federal grants, for the system as a whole, is the direct rewarding of that activity and the penalizing of the opposite activity (plans 3 or 4).

Perhaps the most interesting point to notice is that a general increase in federal support over the control plan may reduce total state expenditures; this is what one observes in moving from plan 1 to plan 2, where the numerator of Z shrinks and its denominator enlarges. What was intended to stimulate may in fact retard. If this proposition is valid, then its explanation may be sought in the expectations held by state administrators concerning future flows from external sources, a factor to which we only alluded earlier and upon which we would encourage research.

As a closing comment, we would add that, although the preliminary quality of the simulations has already been mentioned, the results associated with plans 1 and 2 are not as tightly tethered by this qualification as are the other plans.

Footnotes

*This research was supported by the National Science Foundation Grant GS-1926.

[1]All data are taken from (2), and are defined in some detail there. Minor amendments were made in series whose composition had changed definitionally over time.

References

1. Benson, Charles S. The Economics of Public Education. Boston: Houghton Mifflin, 1961.
2. Bureau of the Census, Compendium of State Government Finances in 19XX (1950 through 1965).
3. Crecine, John P. Governmental Problem Solving: A Computer Simulation of Municipal Budgeting. Chicago: Rand McNally, 1969.
4. Davis, Otto A., M. A. H. Dempster, and Aaron Wildavsky, "A Theory of the Budgetary Process," American Political Science Review, LX (September, 1966), 529-547.
5. Gerwin, Donald. A Process Model of Budgeting in a Public Organization. Doctoral dissertation, Carnegie-Mellon University, 1967.
6. Gerwin, Donald. Budgeting Public Funds. Madison: University of Wisconsin Press, 1969.
7. Hammann, Arthur and Aaron Wildavsky. "Comprehensive Versus Incremental Budgeting in the Department of Agriculture," Administrative Science Quarterly, X (December, 1965), 321-346.
8. Harlow, Robert L. "Sharkansky on State Expenditures: A Comment," National Tax Journal, XXI (June, 1968), 215-216.
9. Miner, Jerry. Social and Economic Factors in Spending for Public Education. Syracuse: Syracuse University Press, 1963.
10. Naylor, Thomas H., Kenneth Wertz, and Thomas H. Wonnacott. "Methods for Analyzing Data

from Computer Simulation Experiments," Communications of the ACM, X(November, 1967), 703-710.

11. Naylor, Thomas H., Kenneth Wertz, and Thomas H. Wonnacott. "Some Methods for Evaluating the Effects of Economic Policies Using Simulation Experiments," Review of the International Statistical Institute, XXXVI (1968), 184-200.

12. Norman, M. R. "OLS Autoregressive Program," Economics Research Unit, University of Pennsylvania.

13. Osman, Jack. "On the Use of Intergovernmental Aid as an Expenditure Determinant," National Tax Journal, XXI (September, 19 1968), 437-447.

14. Sharkansky, Ira. "Some More Thoughts about the Determinants of Government Expenditures," National Tax Journal, XX (June, 1967), 171-177.

16. Walker, Jack L. "The Adoption of Innovations by the American States," unpublished manuscript, University of Michigan.

17. Wildavsky, Aaron. The Politics of the Budgetary Process. Boston: Little, Brown, and Company, 1964.

THE USE OF SIMULATION TECHNIQUES TO STUDY THE CHANGING ECONOMIC SITUATION OF THE AGED

James H. Schulz, University of New Hampshire

Introduction

Due to inability or neglect, a great many people in the United States have failed in recent years to prepare, through individual or group action, for an ever growing number of retirement years. This has resulted in the plight of today's poverty-stricken retired aged population. Enough reliable statistical information has now been accumulated to substantiate clearly the existence of a relatively low income status for all but a very small minority of the current retired aged in the United States. Using the Social Security Administration's highly stringent definition of poverty, for example, three out of every ten people 65 and older -- in contrast to one in nine younger people -- were living below the poverty line in 1966. Another one-tenth of the aged population was on the border of poverty.

Recently the U.S. Senate's Special Committee on Aging undertook a year-long investigation of the economic problems of the aged. A working paper [1] prepared for the Committee by a Task Force of persons specializing in research in this area made the following major points concerning the current situation:

1. Americans living in retirement are suffering from an income gap in relation to younger people. And as the gap widens, low income continues to be the number one problem facing most of our 20 million persons 65 years or older, as well as other millions just a few years younger.

2. More Americans are spending more years in retirement periods of indeterminate length and uncertain needs, causing a mounting strain on resources they had when they began retirement. For an ever-rising proportion of women -- most of them widows -- the problem is especially severe.

3. Today's inadequacies in retirement income -- and the policies and trends that perpetuate them -- should be of direct concern not only to our population of aged and aging Americans, but also to those in middle age or younger. Most parents today face a common problem: How can they allocate earnings to meet current obligations to their family and still have something left over for retirement?

4. Projections and various studies indicate that Social Security, private pensions, and other forms of retirement income are not improving fast enough to reverse or significantly counter present economic trends.

5. Facing what must be recognized as a worsening retirement income crisis, the Nation must take positive, comprehensive actions going far beyond those taken within recent years. The Nation faces these basic policy issues:
a) What is an adequate level of income for retired persons?
b) What part in attaining this level should be played by government programs, by voluntary group action, and by individual effort?
c) Is the economic problem of aging a temporary problem that requires a different solution or a different "mix" of solutions for today's aged than for those reaching old age in the future?

Whether or not one agrees with the previous summary of the current situation, it seems clear that there is a need for research which focuses on the unique economic problems of the rapidly growing group of retired aged families -- families distinguished not just by the older age of their members but also by the fact that none of their members, including the head, contributes to the family's support by working in the labor force.

For most families, the income problem in old age grows out of the cessation of earnings of one or more family members and the failure of private savings and/or private and public pensions to replace a sufficiently large proportion of these earnings. For a significant number of retirees, however, very low incomes were a problem throughout worklife; some form of income maintenance was in order even before they became too old to work.

Income-maintenance issues as they relate specifically to the elderly require further research into (a) questions of the present and future income and wealth levels in old age; (b) the relationship of these levels both to pre-retirement income levels, to general budgetary requirements of the elderly, and to the income-needs gap of particular aged groups; (c) the transitional period into retirement and the decision-making process involved; and (d) considerations of the adequacy of current income maintenance institutions and the costs of alternative financing arrangements for these institutions.[2]

Simulations of the Aging Process

In order to investigate various aspects of these issues, a "life process" simulation model has been constructed to permit those activities of individuals to be simulated which have an important influence on their economic situation in retirement.* These activities can be divided into the following four categories:

a) Demographic.
b) Work force and earnings.
c) Pension status.
d) Asset accumulation.

A large sample of 33,280 persons in the U.S. population, who were, in general, between the ages of 45 and 60 in 1960, is aged 20 years, using the simulation process. The basic data used are from the "one-in-a-thousand sample," a set of tapes produced by the U.S. Bureau of Census which contains separate records (including demographic, work force, and income information) of a 0.1 percent sample of the U.S. population as recorded in the 1960 census.

At the end of 20 years, these people are age 65 or over and represent the aged population in 1980. Naturally not everyone in 1960 between 45 and 60 can be expected to live at least 20 years. Hence the first life process activity considered in the simulation model is death. A probability of death for each particular year is specified for individuals based on their sex, race, and age. A random drawing from the associated probability distribution is used to determine whether an individual will die or live that year. Similarly, probabilities are specified for other possible occurrences built into the model -- labor force exit and entry, job change, pension coverage, vesting, periods of unemployment, etc.

Each possible "occurrence" specified in the model is treated in a manner similar to the live-die occurrence -- each person being considered in turn. By sequential handling of the various occurrences it is possible to make the consideration of any one occurence dependent on occurrences which had been handled before it. Once one year's simulation is completed, the individual, if he has survived, is aged another year and the process immediately repeated. This continues until the year 1980 is reached (i.e., completion of 20 "passes" in the computer). After all individuals have been processed, the resulting sample population represents the major part of the future aged population, since the surviving individuals are now 65 to 85 years of age.

During the simulation, work income, pension coverage, and asset histories are kept for each individual. Social security benefits for those persons retired and no longer in the work force (and eligible) can be calculated by applying the average "creditable" wage income generated by the simulation to a social security benefit formula. Where applicable, private pension benefits and government pensions can be estimated based on the wages and/or years of service of employees.

Having calculated private and public pension benefits, a "census" is

* A detailed explanation of the simulation model and assumptions are contained in [3].

taken of the retired population at the end of the simulation run, and various distributions of pension income and assets for couples and unrelated persons can be derived.

Previous Findings

The results of the initial simulation study (which have been reported elsewhere) were definitely not encouraging with regard to the future economic situation of the retired U.S. population in 1980. The study [3] concluded that if pension systems are to be used to eliminate poverty among retired families and individuals in the United States and also to improve the relative economic status of the retired population, significant changes in present U.S. pension systems must take place.

Table 1 contrasts the resulting pension income distributions for couples, for example, using two different assumptions regarding the trend of social security benefits. Estimate I assumes a 4 percent annual increase in social security benefits, a 3 percent annual increase in private pension benefits, and periodic increases in "maximum creditable earnings" for social security purposes.

Estimate II simulates a situation in which Congress and the President in, say, 1970 attempt to improve the position of the retired aged relative to that of the working population by increasing benefits 50 percent in that year. Thereafter benefit levels are increased by 2 percent annually to take account of general price increases.

There is a substantial upward shift in the income distribution for Estimate II. The proportion of couples with pension income over $5,000, for example, increases from 12 to 20 percent. The estimate, however, also shows the difficulty of improving the income situation of low-income couples by across-the-board percentage increases in social security benefits. Despite the very large pension increase assumed for Estimate II, almost two-fifths (39 percent) of the retired couples are projected as having total pension income of less than $3,000 in 1980.

Another simulation study which I have conducted [4] measured for the first time the "adequacy" of U.S. pension systems by comparing projected pension income with average earnings of retired persons for the 1, 5, and 10 year period prior to each individual's time of retirement. This simulation study projected pension-earnings ratios for persons retiring in the United States between 1960 and 1980. The projections indicated that U.S. pension systems as they are presently developing are failing to generate for large numbers of aged persons retirement income sufficient to meet generally accepted international and national standards regarding the relationship which pension income should bear to preretirement earnings.

Comparative Study of Social Security Systems -- West Germany and the U.S.

In recent years we have seen developed and implemented in various industrialized countries a number of highly innovated social security systems. These new systems were in large part motivated by dissatisfaction with the existing programs of old age income maintenance in each country. Innovative public pension developments in West Germany, Austria, and Sweden, for example, have been watched by social security researchers and policy-makers with increasing interest.

Given the continuing interest of countries in evaluating and improving their pension systems, study of alternative pension systems is a necessary part of the information required to formulate public policy in this area.

A comparative simulation study of the U.S. and West German social security systems is currently underway. The West German system is being evaluated with regard to the desirability and feasibility for adopting a somewhat similar system in the U.S. For this study, the West German system has been programmed for computer simulation analysis and its major provisions have been adopted to be compatible with the basic provisions of the existing social security retirement system in the United States. Using U.S. data on persons approaching retirement,

Table 1. Projected pension income distribution for retired couples, assuming different social security benefit increases, 1980. (Percentage distribution)

Pension income	Social security income[2] I 4% annual social security increase	Social security income[2] II 6% annual social security increase	Total pension income[1] I 4% annual social security increase	Total pension income[1] II 6% annual social security increase
Total percent	100	100	100	100
Under $1,000	4	0	5[3]	1
$1,000-1,999	21	16	16	12
$2,000-2,999	43	39	28	26
$3,000-3,999	30	27	25	23
$4,000-4,999	2	16	14	18
$5,000 and over	0	2	12	20

1. Pension income includes benefits from social security, private pensions (including State and local government plans), and Federal retirement programs.
2. Social security recipients only.
3. Includes couples with no pension income.

Source: Simulation study. Table reproduced from The Economic Status of The Retired Aged in 1980, Research Report No.24, Social Security Administration (1968).

estimates of aggregate pension income, pension income distributions, program costs, and the effects on poverty groups are being made. These estimates can then be compared with alternative estimates which assume that the current U.S. system does not change radically in the future.

A major principle upon which the current West German social insurance system is based is that the worker should receive a pension which will enable him to achieve a retirement standard of living similar to the living standard maintained during his working life. The retirement pension is related, however, not just to prior earnings of the worker himself (as in the U.S.) but also to the average earnings of all workers at the time a particular worker retires.

The result of these provisions is to provide substantial pensions to workers with long years of service -- substantial in the sense that the pension is a high proportion of average preretirement earnings. Currently a West German worker who has 40 years of service in the program would receive a pension which was equal to 60 percent of his average preretirement earnings. If he has 50 years of service, he would receive 75 percent.

In order to simulate the effect of such a system on pension incomes in the United States, a number of working rules are specified. These are the following:

1. Only persons over age 61 are eligible for a pension.

2. If the female member of a family is not eligible for a pension in her own right, she receives no pension (i.e., no wife benefits are awarded).

3. Pensions are based upon average earnings for a 10-year period prior to retirement (rather than as in the German system -- lifetime average earnings).

4. Workers eligible for social security pensions are assumed to enter the work force at age 20.

5. Pensions are increased 2.5 percent each year after retirement to adjust for general price level increases.

Table 2. Projected total pension income distribution for U.S. retired couples, 1980.
(Percentage distribution)

Total pension income 1	U.S.system 6% annual increase	German system with 0.111 constant
Total percent	100	100
Under $2,000 2	13	16
$2,000-2,999	26	10
$3,000-3,999	23	14
$4,000-4,999	18	12
$5,000-5,999	18	37
$6,000 and over	2	11

1. Pension income includes benefits from social security, private pensions (including State and local government plans), and Federal retirement programs.
2. Includes couples with no pension income.

Source: Simulation Projection.

Table 2 presents the first findings from the study described above. The U.S. system projections are the same as those cited in Table 1; they assume a 4 percent annual social security increase. Table 2 shows a dramatic improvement in the projected pension income distribution when the West German system is substituted for the U.S. one. Almost two-thirds of the couples are projected as having pensions of $5,000 or more.

To some, the pension incomes indicated by the West German system projections may seem, at first glance, to be unrealistically high. However, one should remember, first, that the German social security system is designed to produce pensions which are a high proportion of preretirement earnings and, second, that average gross earnings in U.S. manufacturing currently are about $6,000 a year and growing over time. Thus, Tables 1 and 2 demonstrate a basic fact: the present U.S. social security system, unless changed in a very substantial way, will not (together with private pensions) produce incomes which would permit U.S. workers to maintain their standard of living as they move into the retirement period. Serious consideration should be given to alternative systems which do.

References

[1] U.S. Senate Special Committee on Aging, Economics of Aging: Toward a Full Share in Abundance, 91st Congress, 1st session, U.S. Government Printing Office, Washington D.C., 1969.

[2] Schulz, James H. and Juanita M. Kreps, "Economics of Aging," in Gerontology Society, Research and Development in Social Gerontology, Report to the U.S. Administration on Aging (mimeograph), 1969.

[3] Schulz, James H., The Economic Status of the Retired Aged in 1980: Simulation Projections, Social Security Administration Research Report No.24, U.S. Government Printin Office, Washington D.C., 1968.

[4] Schulz, James H., "Aged Retirement Income Adequacy -- Simulation Projections of Pension-Earnings Ratios," in U.S. Joint Economic Committee, Old Age Income Assurance, Part II, U.S. Government Printing Office, Washington D.C., 1968.

DISCUSSION

Lindsey Churchill, Russell Sage Foundation

Introduction

I came to the field of computer simulation very recently, in the last two years. I am a sociologist by training, and as you undoubtedly know, sociology is just beginning to use computer simulation in its problems. Therefore, I am a novice in this area. However, I have had a good deal of experience with mathematical models and statistics in sociology, and I discovered that if I thought of a simulation man as a mathematical modeler with a powerful friend, the computer, I didn't go far wrong. I found that I could bring an evaluation scheme that I had used on mathematical models over to the study of computer simulation, and I will evaluate the present papers in terms of it.

My scheme consists of evaluating a computer simulation in three areas:

a. Degree of understanding of the substantive problem;

b. Amount and sophistication of the statistical testing of the simulation; and

c. Amount and sophistication of the mathematics used in the simulation. (This scheme is not original with me, I am sure.)

I turn first to the paper by Manser, Naylor, and Wertz1/, then to Schulz' paper 2/, and finally to a comparison of the two. In advance, let me say that I like both papers. I will make a number of critical comments, but I intend them only to place the papers in perspective. I am sure that the authors are aware of my criticisms, but simply couldn't do everything at once.

Evaluation of Manser, Naylor, and Wertz

The first question to ask is this: Exactly what are the authors trying to do in this paper? They are trying to invent a strategy by which Federal policy makers can disburse aid to education at the State level so as to "stimulate" the States to pay out the greatest sums possible for education. That is, they want to find a way to manipulate the purse strings for education in each State. In particular, they would like to prevent States from cutting down on their own expenditures by using Federal funds to make up the difference.

How do they proceed to find such a strategy? Very sensibly, they start from the point of view of the State budget-maker for education. They argue that he takes three major variables into consideration: the size of last year's education budget; the sizes of last year's education budgets in neighboring States; and the size of last year's Federal grant to the State for education. Further, the same three factors are assumed to be the central ones in each State.

They show that a linear weighting of the three variables obtained through a regression method rather accurately reproduces the budgets for each State for the years 1951 through 1965. Since the regression function is a good predictor of budget size, the authors argue that they can study the effects of possible Federal strategies in giving education money by working with the function rather than by trying the strategies out in practice.

The authors then try out six different strategies that Federal people might use to induce State budget-makers to increase their education budgets. They project budget sizes for each strategy for the years 1966 through 1972, and use their statistic Z as the criterion for the relative success of the strategies. Their conclusion is that the two strategies that penalize States for using Federal money in place of State money and that reward States for using Federal money in addition to State money maximize the total amount spent collectively by the States for education. (The penalty used was a smaller amount of Federal aid than usual, while the reward was a larger amount of Federal aid than usual.)

Somehow, I don't find this result surprising. I can readily understand that State budget-makers would be responsive to a "carrot-and-stick" strategy. On the other hand, I am surprised that the authors found this classic strategy with their method. I don't see how the "static" model underlying the authors' regression scheme allows a "dynamic" conclusion like this one.

Exactly how do the authors reach this conclusion? I think they make an inferential leap from their statistics to a theory of budget-making that is unwarranted. (Actually, the authors agree about the inferential leap. They say that their results only suggest the conclusion. However, the leap is vital to their paper since they are after a strategy to loosen State strings, and it is important to see how they did it.)

Using their regression method, they found the best set of coefficients to typify each State over the entire period 1951-1965. They interpreted the coefficients c^i of the Federal aid variable as indicators of the "stimulation" or "constriction" effect of Federal aid. Coefficients less than one indicate the "constriction" effect, or substitution of Federal aid for State money; coefficients greater than one indicate the "stimulation" effect. Thus, States are fixed in advance as "stimulated" or "constricted" by the nature of the regression method; it only remains to find out which state a State is in.

It is not obvious that a "carrot-and-stick" strategy should emerge as superior because there is no provision for the States to react to the strategies. The regression method here does not contain a simulated way by which States can respond to Federal policy over a period of time. Why, then, do Plans 3 and 4, the "carrot-and-stick" strategies, emerge as the best? The answer seems to be that the arithmetic of the calculation of the statistic Z just comes out that way.

The statistic Z can be expressed as follows:

$$\Sigma\Sigma S_t^i P_t^i \;/\; \Sigma\Sigma G_t^i P_t^i =$$

$$\Sigma\Sigma [a^i S_{t-1}^i + b^i N_{t-1}^i + c^i G_{t-1}^i] P_t^i \;/\; \Sigma\Sigma G_t^i P_t^i \quad .$$

In this expression the S^i , N^i , and P^i are all positive and the same for each of the six strategies being compared. The a^i and b^i are almost all positive (only eight small negative values among 96) and the same for all strategies. The c^i are the same for all strategies, but many of them are highly negative. The only varying quantities in the different strategies are the G^i , the amounts of Federal aid to the different States.

The "carrot-and-stick" strategies allow G^i to be smaller than in the other plans if the corresponding c^i are less than one (in Plan 3) or less than zero (in Plan 4). When the c^i are negative, the third term in the numerator of Z is negative and reduces the size of the numerator, tending to reduce the size of Z. But Z will be larger the smaller the amount of reduction due to this term. And the third term will tend to be least negative in the two "carrot-and-stick" strategies because the minimum G^i's are combining with the negative c^i's. Further, the minimum G^i's will make the denominator of Z a minimum, thereby contributing to making Z a maximum for these strategies.

How, then, do the authors get from this arithmetic result to the inference that a "carrot-and-stick" strategy is superior to other strategies? I think the authors may have proceeded in the following way. They discovered that the multiple R's for the regression equations were very high (the median R is 0.96). Having made this discovery, the authors turned the regression function into a theory of how budget-making gets done. There were many different theories that would be consistent with the regression functions, but the one that the authors implicitly chose seemed to be the following one.

In making this year's budget a State budget-maker first adds in an amount equal to the size of last year's budget. He (ordinarily) adds an increment to that. He then (ordinarily) adds a second increment to the amount so far cumulated, based on the sizes of last year's budgets in neighboring States. Then he looks at the amount of Federal aid to his State for last year. He expects to get the same amount of Federal aid this year, or possibly more. Now he is at a critical choice point: he can either add in a third amount to his cumulating total that is larger than the amount of last year's Federal aid; or he can add in a third amount smaller than that sum; or he can subtract an amount based on that sum of Federal money. If he takes the first option, he is said to be "stimulated" by the Federal money, because he adds in all the anticipated Federal money in this year's budget and more besides, which must come from the State. If he takes the second or third option, he is said to be "constricted" by the Federal money, for the following reasons. In the second option he adds in a smaller amount of anticipated Federal money than he expects to receive, thereby expecting to have some Federal aid left over to substitute for State money already budgeted. In the third option he actually reduces the size of the total cumulated on the basis of the first two variables, but since he expects to get the Federal aid anyway, all of it can be substituted for State money that would otherwise have to be spent.

While it is true, of course, that budget-making gets done somehow, the authors present no evidence to show that it gets done in the way outlined above. Until that evidence is collected, the authors' conclusion must remain an unwarranted inference. The most that their analysis allows them to say is that the best-fitted linear combination of the three variables has a very high correlation with the size of the education budget.

Suppose the authors could present evidence that supported the theory discussed above. A simulation could then be made of the theory. The six strategies could be compared much as the authors do here, and it is entirely possible that the "carrot-and-stick" strategies would emerge as superior to the others. The conclusion that the authors would like to draw would have more weight than it does at present because it would be based on some kind of theory of budget-making. The accuracy of the conclusion would rest on the degree to which the theory of budget-making captured the actual budget-making process.

Thus, in terms of the evaluation scheme I outlined earlier, the authors have made some progress in the understanding of the substantive problem, but not a great deal. The high multiple R's show that the three variables selected are somehow related to the process of budget-making, but that is as much as can be safely inferred. Further understanding rests on evidence collected by other methods of study than the present one, methods that try to find out directly how budget-makers proceed and why.

But a deeper level of understanding may not be necessary to provide a good practical answer to the question that the authors wish to answer: How should the Federal government spend money for education so as to stimulate the states to spend the greatest amount? I think the authors' "carrot-and-stick" proposal is a good one to try, even though their evidence doesn't support it.

In terms of amount and sophistication of statistical testing, the authors do very well. They are aware of the problems involved in doing regression on time series. They use a sophisticated method for constructing simultaneous confidence intervals to compare the different plans. Best of all, they replicate the simulation of each proposed plan thirty times, something not done too often in simulation work.

Finally, in terms of mathematics, I find very little mathematics done. The paper rests entirely on known statistical techniques. There is some minor mathematics involved in adapting the regression function to the different strategies for Federal spending, but none at all on the substantive issues of how budget-makers make budgets.

In summary, I think the authors have convincingly demonstrated the relevance of their three variables in the budget-making process. Beyond that, I am not convinced by their analysis that the "carrot-and-stick" strategy is the best one for Federal people to adopt, yet on common-sense grounds I find it a perfectly plausible strategy.

I turn now to Schulz' paper.

Evaluation of Schulz

As has been mentioned, Dr. Schulz was originally scheduled to be a discussant in this section. At the last minute he was pressed into service to give a paper. The paper he just gave is an outgrowth of research described in his monograph, The Economic Status of the Retired Aged in 1980: Simulation Projections.[3] My discussion is based on that monograph since I did not have a copy of his paper in advance.

Let me begin by asking the question: Exactly what did Dr. Schulz try to do in this research? He answers in the following way:

> "Specifically, the study projects the pension income and assets of retired persons in 1980 and investigates the role of private pensions and the U.S. social insurance system in providing retirement income." (p. 1)

The problem is similar in form to the problem pursued by Manser et al. Dr. Schulz seeks to estimate the income of a retired person in the future, while Manser et al. seek to estimate a State's expenditure for education in the future.

How does Dr. Schulz proceed? He proceeds by carefully delineating the major features of American society that affect the size of a retired person's income. These include age, sex, race, marital status, present income, employment status, and private pension plan status, among others.

The projections to 1980 were carried out by selecting all married couples where the husband was between 45 and 60, and all unmarried persons between 45 and 60, from the 1/1,000 Census sample of 1960. These persons were then "aged" into the future year by year on the basis of the simulation model.

A variety of things can happen to a person in a twenty-year period, and Dr. Schulz had to take the most important things affecting income into account. Persons can die in the given time period; they can leave the work force, re-enter it, leave it again, etc.; they can change jobs; they can retire early; they may or may not be covered by a private pension system; they may accumulate financial assets, such as equity in a home; they spend varying lengths of time in each job; etc.

The mechanism of the "aging" process is stochastic. The probability of death in a given year is estimated, as are the probabilities of changing jobs, leaving the work force, being covered by a private pension plan, etc. Then the probabilities (not necessarily assumed constant for each year) are applied to the persons in the 1/1,000 sample for each year from 1961 to 1980. Persons who "die" drop out; persons who "live" are surveyed in 1980. The incomes and assets of persons fully "retired" by the simulation by 1980 are then tabulated to form the basic output of the simulation.

Essentially then, each person selected from the 1/1,000 sample is assigned a history for the period 1961-1980 so that his income and assets may be estimated for the year 1980. The great power of simulation can then be put into play by rerunning the simulation under different assumptions about social security benefits, about the degree of vesting in private pension programs, about the rate of early retirement, etc., to see what would happen to the income distribution if these assumptions were put into practice before 1980.

The general conclusion reached by Dr. Schulz is that while there will be improvement in the retirement income picture by 1980, fifteen percent of retired couples would have total money income (pension plus asset income) less than $2,000, and 35% would have total money income less than $3,000. Analogously, seventeen percent of retired unmarried persons would have total money incomes less than $2,000 and 41% would have total money income less than $3,000. I find this conclusion disheartening.

In terms of understanding the substantive problem, I think Dr. Schulz has done an excellent job. It would have been easy for him to reduce the elaboration of his simulation by simply ignoring many refinements. Dr. Schulz created a much larger number of different types of future history than many others would have, and I commend him for his painstaking efforts to be accurate.

In terms of statistical testing, however, Dr. Schulz is much sketchier. He made only one simulation run under a given set of assumptions, and he made only eight simulation runs in all. His results would be more convincing if they were averages from many identical simulation runs.

Finally, in terms of my evaluation scheme, Dr. Schulz uses mathematics very little, mainly to estimate changing probabilities as the years move from 1961 to 1980. He avoids the use of extensive mathematics by baldly assuming that persons' future histories are independent of their past histories. He doesn't worry about whether or not he gets the right history attached to the right person; he only tries to ensure that the right numbers of events in different categories will occur.

In summary, I think that Dr. Schulz has made a convincing case that the income for retired persons in 1980 will be better than it is now, but that large numbers of persons will still be below the present poverty line. The conclusion would be more convincing if the simulation runs had been replicated a number of times.

I turn now to a comparison of the two papers and some general comments.

Comparison of the two papers

In terms of my evaluation scheme Manser et al. and Schulz complement each other in the first two areas. Manser et al. are strong on statistical testing and weaker on substantive understanding. Schulz is just the opposite, being strong on substantive understanding and weaker on statistical testing. On balance, I think I prefer Dr. Schulz' use of the available time. I prefer to spend more time on understanding the substantive problem than on statistical tests of whatever understanding I have, but that is purely a personal preference. I fully recognize the importance of statistical testing.

However, it is the comparison in the third of my areas, the use of mathematics, that leads to what I think is the heart of present difficulties with many simulation studies. Neither paper uses mathematics in more than a minor way to deal with substantive issues. I take the relative lack of mathematics in the present studies to indicate that we do not yet have the requisite substantive understanding necessary to make thoroughly convincing simulations. For example, it seems to me that a strong theory of budget-making, in Manser et al.'s case, or of accumulating personal savings, in Schulz' case, would consist of mathematizations of psychological, sociological, and economic findings about these problems. Looking ahead, I expect that future simulation studies on the present problems will "reek" with mathematics.

Dr. Schulz' paper is instructive in this regard because he spent a great deal of time laying out the substantive issues in the income of retired persons. But at a critical point he brought in the most simplistic of mathematical assumptions, statistical independence, by randomly assigning future histories to persons. (I say "simplistic" only in reference to a strong theory of how persons' futures are tied to their pasts.) I think Dr. Schulz was justified in making this assumption because his task was basically practical and not theoretical: what will the income distribution look like in 1980 closely enough so that certain gross conclusions may be drawn correctly?

I urge, therefore - as discussants have urged from time immemorial - that the authors not lose sight of our deep theoretical problems. It would be easy for Manser et al. to "export" their regression approach to other problems without advancing our theoretical understanding of the substantive issues involved; similarly for Dr. Schulz. And these studies would have practical importance of the kind seen here. Nevertheless, I urge our authors not to take this easy road but to devote some of their talent and energy to the very difficult substantive problems involved. I urge our authors in this way because I believe that essentially new advances in simulation rest on advances in understanding these deep problems.

In conclusion, I want to praise the authors for actually having completed their simulation studies. I have discovered that a lot of enthusiastic talk goes on about the possibilities of simulation, but that relatively few studies get done. I am delighted that the present authors have escaped that easy fate.

1/ Marilyn E. Manser, Thomas H. Naylor, and Kenneth L. Wertz, "Effects of Alternative Policies for Allocating Federal Aid for Education to the States: A Preliminary Analysis," Econometric System Simulation Program, Working Paper No. 34, May 15, 1969.

2/ James H. Schulz, "The Use of Simulation Techniques to Study the Changing Economic Situation of the Aged," August 1969.

3/ James H. Schulz, The Economic Status of the Retired Aged in 1980: Simulation Projections, Research Report No. 24, Office of Research and Statistics, Social Security Administration, U.S. Department of Health, Education, and Welfare, Washington: U.S. Government Printing Office, 1968.

XI

CONTRIBUTED PAPERS I

Chairman, ABBOTT L. FERRISS, Russell Sage Foundation

ON ANALYZING AND PREDICTING ENROLLMENTS AND COSTS IN HIGHER EDUCATION

Gus W. Haggstrom, University of California at Berkeley

This paper is a preliminary report on a study of enrollment and financial statistics in higher education that has been sponsored in part by the Carnegie Commission on Higher Education. The purpose of the study is to analyze statistically certain trends in higher education and to assess the effects of factors such as the Vietnam war, the G.I. Bills, and the increased holding power of the high schools. The statistical models which have been used in analyzing the data are described in detail and compared with other models which have been used in the past. This report also contains a discussion of current trends in higher education, some predictions of future demands on higher education, and some long-term projections of enrollments and costs.

The colleges and universities have grown at an almost incredible rate throughout the last decade. This growth is attributable in part to the growth of the college-going age group as an aftermath of the post-World War II baby boom; in particular, the population in the 18-21 age group has risen by almost 60% from 1958 to 1968. During the same ten-year period the high schools have cut their drop-out rate; over 75% of the young people are graduating from high school today as compared with only 65% ten years ago. Also, an increasing proportion of the high school graduates go on to college--approximately 60% at the present time as compared with slightly over 50% only ten years ago.

As a result enrollments in colleges and universities have soared. Despite the negative effects of the Vietnam war, undergraduate enrollment has more than doubled during the last decade. Graduate enrollments have risen even more rapidly, doubling in the last 7 years, tripling in only 12 years. The total opening fall degree-credit enrollment in institutions of higher education in 1968 was 6.9 million students as compared with 3.2 million only ten years ago, an increase of over 110% in ten years.

With the high schools continually reducing their drop-out rates and with an increasingly higher percentage of high school graduates going on to college, institutions of higher education are being asked to serve an ever-broadening segment of the population. Nevertheless, the yearly retention rates for college students have remained virtually unchanged at a surprisingly high level for the last 20 years, contradicting the popular notion that college work is so demanding that only the most able survive the first year.

These conclusions suggest that on the whole the higher educational system in this country is continually adjusting to accommodate an increasingly larger segment of the population. The notion that a college education is something to be reserved for only the more academically talented students appears to be eroding with time in almost the same way that the corresponding notion about high school education did 30 or 40 years ago.

Before the methodology used in this study is discussed, let us turn to some projections of future demands to be placed upon higher education. The table below shows how much time the institutions of higher education will have to increase their capacities by 50% (100%) in various categories according to the projections given in this paper. It also shows how long it has taken to increase their capacities by 50% (100%) in the past.

According to the projections given in this report, total opening fall degree-credit enrollment will continue to rise at a rate of about one-half million students per year for the next ten years. Then total enrollment should begin to level off and remain rather stable throughout the 1980's at around 12½ million students with a very slight decline beginning perhaps in 1982 or 1983.

No one seems to doubt that the demands upon higher education will continue to rise at a rapid rate for some time, but there is considerable reason to doubt that colleges and universities can continue to meet these demands as they have in the past. In particular, with the current financial plight of the private institutions due to spiraling costs and tuition rates, these institutions are apparently being priced out of the market and may fast be approaching the limit of their capabilities insofar as further increases in enrollment are concerned. The public institutions are also having to cut back on some of their programs as state legislators and governors react to rocketing costs and student disruptions on campus with substantial budget cuts. Also, recent cuts in research support by various government agencies may already have seriously affected future growth at the graduate level. Nevertheless, on the basis of past performance, it's hard to discount the capability of the colleges and universities to meet these demands, despite the mounting financial crisis in higher education.

As a partial indication of the methodology used

	Projected Number of Years to Increase by 50% (100%)	Number of Years Needed in the Past	1968 Level
Undergraduate enrollment	9 (40)	5 (9)	6,131,000
Graduate enrollment	4 (7)	5 (7)	797,000
Total enrollment	8 (26)	5 (9)	6,928,000
Bachelor's and first-professional degrees	7 (27)	5 (12)	667,000
Master's degrees	4 (7)	3 (6)	177,000
Doctor's degrees	5 (8)	4 (7)	23,000
Expenditures	4 (8)	4 (6)	$20.6 billion

in this study, the analysis of the undergraduate enrollment data will be briefly discussed. In assessing the growth of undergraduate enrollment, it seemed reasonable to compare the number of undergraduate students in each year with the number of high school graduates over the past several years. This led us to consider the "undergraduate enrollment rates"

$$r_t = U_t / \Sigma_{k=0}^{3} H_{t-k}$$

for each sex, where U_t denotes the opening fall undergraduate degree-credit enrollment during year t and H_t is the number of high school graduates during year t. These enrollment rates for women have been increasing consistently since World War II at a rate of about 1% per year, whereas the enrollment rates for men have fluctuated in a rather wild manner, with dips coinciding with high draft calls and bulges with the return of large numbers of veterans to civilian life. The basic assumption in analyzing male undergraduate enrollment is that, if one first accounts for the wartime effects upon undergraduate enrollment, the enrollment rates r_t for men can be suitably approximated by a logistic growth curve

$$\rho(t) = \gamma/[1 + e^{-(\alpha+\beta t)}].$$

This led us to use the nonlinear regression model

$$U_t=\rho(t)\Sigma_{k=0}^{3}H_{t-k}+\beta_2 VII_t+\beta_3 VK_t+\beta_4 VV_t+\beta_5 DK_t+\beta_6 DV_t+e_t$$

where VII_t, VK_t, and VV_t denote the numbers of veterans attending institutions of higher education from World War II, the Korean War, and the Vietnam conflict respectively; and DK_t and DV_t denote the number of draftees in service for the Korean and Vietnam conflicts respectively. The analysis of the data for the years 1947-68 yields estimates of the underlying growth rate $\rho(t)$ as well as assessments of the effects of the wars upon enrollments. The projections of undergraduate enrollment are obtained by combining extrapolations of the estimated growth curve with projections of high school graduates and estimates of the post-Vietnam effects upon enrollment based upon the Korean War experience.

A DEVELOPMENTAL STUDY FOR AN INDEX OF UNDEREMPLOYMENT

B. V. Shah, Research Triangle Institute

1. Summary. A nation's human resources may be underutilized not only because of unemployment in the usual sense but also because of what economists term "underemployment". Underemployment can be defined as the employment of persons at jobs that call for less than their highest current level of skill and at wages less than those to which their skills, if fully utilized, would normally entitle them. The earnings of an individual when he is employed at his highest level of skill may be called his "potential" earnings. The potential earnings of any individual is not observable, so one could put forth arguments for or against any particular definition of potential earnings. The absolute level of potential earnings will depend in part on adjustments made in current earnings for noncompetitiveness and for the structural-frictional aspects of the economy under consideration.

Basic to the methodology is a definition of a standard of potential earnings that is specific for subsets of the population, rather than for each individual separately; these population subsets are defined by socioeconomic factors such as sex, age, race, education, and region. The subsets with the same combination of these factors but with different occupations form a group. The average potential is the best average earnings the group can achieve and is defined to be the maximum average earnings over all subsets in the group.

Under ideal conditions of no (zero) underemployment, the current average earnings of the group will be equal to the average potential earnings of the group. The differences in the current earnings from the potential earnings provide the basis for an index of underemployment. Two index numbers were developed using two distinct statistical concepts of averages: "median" and "mean". These index numbers were respectively: (1) the number underemployed as a percent of group total, and (2) the loss of potential earnings due to underemployment as a percent of potential earnings.

A preliminary analysis was made using selected data on occupational class, sex, region, race, age, education and earnings from the five percent samples of the 1960 U. S. Census. This analysis indicated that underemployment among nonwhites was worse than among whites, and that the difference between the underemployment index values for nonwhites and whites was larger in the South than in the North and West. From a sample survey of selected North Carolina counties, estimates of these index numbers were obtained to illustrate their usefulness for regional comparisons.

Viewed as techniques for ranking population groups with respect to underemployment rather than as absolute measures of underemployment, the two indexes developed in this pilot study would appear to have considerable usefulness as guides to changes occurring over time and geographical regions. Census year data can be used to estimate the required potential earnings standards and base year indexes. Annual indexes on a regional basis can be calculated using data collected by the Bureau of the Census Current Population Surveys. Current indexes for populations confined to smaller areas such as counties or state economic areas may require surveys of sufficient size to estimate actual earnings and the weights (subset sizes) required by the method.

2. Nature of the Problem. There are a number of attempts to define and to measure underemployment in the literature. The President's Committee to Appraise Employment and Unemployment Statistics, Measuring Employment and Unemployment (Washington: U. S. Government Printing Office, 1962, p. 58) defines underemployment as follows:

"Part of a nation's human resources may be underutilized not because of unemployment in the usual sense but because of what economists term underemployment. Underemployment can be defined as the employment of persons at jobs that call for less than their highest current level of skill (and at wages less than those to which their skills, if fully utilized, would normally entitle them). We shall distinguish this from partial employment or involuntary part-time work, which can be defined as the employment of a person, whether or not at this highest level of skill, for fewer hours per week than he seeks to work. It is also to be distinguished from the losses resulting from the failure, for whatever reason, to train people for the highest skills that their innate abilities would permit them to achieve. Underemployment as here defined is like unemployment in that it results in a loss of income to the individual affected and a loss of output to society."

There have been other attempts to develop theories of underemployment of disguised unemployment as applied to persons whose marginal productivity is zero or negative. The latter term is usually restricted to persons who are not normally engaged in wage employment. The term as defined with reference to marginal productivity is not applied to wage labor, since presumably employers will not employ a laborer for wages unless his labor increases the total production (United Nations, 1951; Libenstein, 1957; Nurske, 1957). In this paper we will not consider the concept of disguised unemployment, but will confine our attention to the concept of underemployment as stated in the report by the President's Committee (1962).

The concept of underutilization of human resources that we are interested in can be explained as follows. If there are many persons seeking a particular type of job for which they are most efficient and the number of such jobs available is less than the number of individuals then some of them may remain unemployed or have only part-time jobs, or some of them may seek other jobs which do not make use of their full potential and consequently their earnings will be smaller; or there may even be persons who would accept similar (or identical) occupations at lower renumeration. Alternatively, there may be individuals who may remain in jobs with lower income in spite of the availability of jobs with higher earnings, for reasons of their own, such as climate, type of work, or leisure time. The underemployment resulting from these free choices corresponds to voluntary underemployment and not to the involuntary underemployment implied when usually referring to underemployment. There does not seem any easy way of separating these two types of underemployment and hence, such individuals will be considered underemployed by the definition.

The absolute level of underemployment will depend on what we define as the 'potential earnings' or 'ideal earnings' of a group of individuals. It is noted that the potential earnings of any individual is not observable and hence one could put forth arguments for or against any particular definition of potential earnings. The absolute level of potential earnings will depend in part on adjustments made in current earnings for noncompetitiveness and for the structural-frictional aspects of the economy under consideration.

It is obvious that the difficulties involved in defining and measuring the absolute level of underemployment are great. The attempt here will be to define a reasonably good relative measure, but no claims will be made for its merits as an absolute measure. In fact, the use of the maximum average earnings over all occupational subsets as the potential earnings for a population group, and also the inclusion of all those unemployed as underemployed, cause the absolute values of the measure to appear very high. However, the measure as defined is expected to be suitable for comparing various areas or regions or the same area or region over different time periods. It is possible that, like any other index number or aggregate measure, there will be biases in the indices that are defined. As time passes there will be a need, in particular, to change the base year of comparison and to otherwise improve the index.

3. Definition of Potential Earnings. The assumptions and postulates that we make in this section will be with respect to a group of individuals and as such will not be valid for each individual of the group. Any attempt to apply these to individuals in the group will be completely meaningless.

Suppose there is a group of individuals with a specified set of characteristics; for example, consider all white males residing in the North region of the United States, aged 24 years and with a high school education or less. Some of these may be clerks in offices, some may be sales workers in department stores, some may be machine operators in factories and others may have selected alternative occupations. Let those individuals of the group having similar jobs form a subset of the group. Data are available to find the median earnings or mean earnings of the individuals in each of the subsets.

In general, we consider a group of individuals with characteristics $x_1, x_2, \dots, x_n$ such as sex, race, age, education, and then consider a subset of this group having a particular occupation with characteristics $y_1, y_2, \dots, y_m$. Let the mean earnings of the individuals in the subset be $\bar{I}(X,Y)$ and the median earnings be $I^*(X,Y)$ where X and Y are vectors representing various characteristics of the individual and his occupation. With reference to the subsets described above we make the following two assumptions:

Assumption 1. The mean potential earnings of the group with characteristics X is equal to the maximum mean earnings over all subsets with the same X but having different occupations Y. Thus,

$$\bar{P}(X) = \max_{Y} \bar{I}(X,Y) \quad .$$

Assumption 2. The potential median earnings of the group with characteristics X is equal to the maximum (median) earnings over all subsets with the same X but having different occupations Y. Thus,

$$P^*(X) = \max_{Y} I^*(X,Y) \ .$$

The assumptions are implied from the fact that the potential is the best that the group "can" achieve and earnings is the criterion. Hence, the potential mean (median) earnings is the maximum mean (median) earnings for the group.

4. Indexes of Underemployment. In the previous section we have defined potential median and mean earnings. The potentials represent ideal levels of earnings for the group. Since existing conditions are different from the ideal, our interest is to measure the extent of departure from the ideal levels. The extent of departure from ideal earnings, for example, will represent the loss of output to society under prevailing conditions.

We define two such measures to represent departure from ideal conditions: (1) the number of individuals underemployed as a percentage (U) of the total number of individuals in the group and (2) the loss of earnings due to underemployment as a percentage (L) of the potential earnings of the group.

Let there be exactly k groups in the population with characteristics vectors $X_1, X_2, \dots, X_k$. Further, let M_i be the number of individuals in the i-th group having earnings less than the potential median earnings $P^*(X_i)$. Under ideal conditions, the actual median earnings of the i-th group would also be $P^*(X_i)$ and we expect that only $(0.5)N_i$ in the group would have earnings below $P^*(X_i)$. Since the present conditions are not ideal, the number $M_i - 0.5N_i$ represents the excess number of individuals having earnings below the potential median earnings for the i-th group. By definition, these are the underemployed individuals. Hence, for the population, summing over groups, we find the number of underemployed is $\Sigma M_i - (0.5) \Sigma N_i$. The percentage underemployed (U) is then:

$$U = \left\{ \frac{\Sigma M_i}{\Sigma N_i} - 0.5 \right\} 100\% .$$

We make use of the potential mean earnings to measure the loss of output to society. Under ideal conditions the mean earnings of all the individuals in the i-th group will be equal to $\overline{P}(X_i)$. Since the actual mean earnings will be equal to $\overline{I}(X_i)$, the loss of output to the society, due to underemployment in this group, is equal to $N_i\{\overline{P}(X_i) - \overline{I}(X_i)\}$. Summing over all groups, we obtain the total loss of output to society as $\Sigma N_i\{\overline{P}(X_i) - \overline{I}(X_i)\}$. Expressing this as a percent of the total potential $\Sigma N_i\overline{P}(X_i)$, we obtain the percentage loss (L) of output to society as

$$L = \left\{ 1 - \frac{\Sigma N_i\overline{I}(X_i)}{\Sigma N_i\overline{P}(X_i)} \right\} 100\%$$

By comparing the measures U and L against the same measures for a standard population in a standard or base year, we can calculate the corresponding indexes of underemployment.

5. Application to United States Data. One of the main questions in the assumptions and measures defined in previous sections is concerned with what constitutes an appropriate set of characteristics X and Y . Any characteristic of an individual likely to have any effect on his earnings may be included in a study. However, there may be practical limitations in that we may not be able to obtain data on some of the characteristics or the cost of obtaining data on other characteristics may be prohibitively large. In some cases an additional characteristic may be highly correlated to the other characteristics already included. By adding such a characteristic we may not be adding much to the information we already have.

We may start our analysis with a limited number of characteristics, such as sex, race, region, age, education, occupational class, on which sufficient data are already available. It would amount to defeatism if we were to take the negative attitude that we cannot make any analysis because we do not have information on all the characteristics considered to have some relevance. If we take an optimistic view, we can argue that a reasonable subset of the relevant characteristics used as independent variables may still provide good estimates of mean or median earnings for the different subsets. There may be biases due to the characteristics that are ignored. However, these biases are likely to be smaller than the "estimates" obtained with the chosen set of characteristics. This will be particularly true in instances where the characteristics are highly correlated. For example, if we consider age and education only and ignore years of experience, in most cases we will not introduce large biases in the mean or the median earnings. In fact, the high multiple correlations in our results for U. S. data in the next section indicate that the characteristics that are considered may be adequate.

It is true that we may need an extensive study to determine the optimum set of characteristics, but not having all the answers on this count does not destroy the validity of the concept or the measurements defined in previous sections.

A limited study was carried out based on the data available from the five percent sample of the 1960 census returns of the United States and also from the data on socioeconomic survey (Benrud, 1968) of some of the counties of the State of North Carolina. The results are presented in the next two sections.

The limitations on the availability of data resulted in using the rather restricted set of characteristics which were included in the Bureau of the Census subject report PC(2)-7B "Occupation by Earnings and Education" of the United States Census of Population 1960. The characteristics X include sex, race, region, age and education. We have analyzed data only for males. The main occupational classes were used as the characteristic Y . The complete details of these characteristics are given in Table I.

Based on the X characteristics, a total of 120 groups were created. Each of these was further divided into ten subsets based on occupational class. The first step was to obtain estimates of mean and median earnings for each of the 1,200 subsets. On studying the data, it was found that mean (median) earnings were not available for a large number of subsets. After omitting the South region and the 18-24 age-group from the data, the remaining 480 cells had very few missing values. Imputed values were substituted for those that were missing by simple interpolation. A preliminary analysis of variance was performed (See Table II). From the results of this study it was concluded that all third and higher order interactions were insignificant. Furthermore, among the second-order interactions, those involving occupation were more significant than the others.

Hence, ten separate multivariate regressions were run, one for each occupational class to represent the effect of various factors on earnings for different occupations. These provided the main effects of all the factors as well as all the second-order interactions involving occupation. The multiple correlation was greater than .95 for most of these regressions and greater than .92 for all ten. The results which seem to provide reasonable estimates are given in Table III. A further check was made by substituting the regression estimates for the cells without any observations and then performing a complete analysis of variance. This analysis was used to derive the estimates of all main effects and interactions. From these regressions, the potential mean and median earnings for each of the 120 groups was estimated by taking the maximum over the ten occupational classes. The potential mean and median earnings along with the occupation class which yields the optimum earnings are presented in Tables IV and V.

Table VI presents three types of earnings:

actual, estimated, and potential earnings. The males in each occupational class were divided into subsets depending on their age, education, race, and region. The subsets were numbered 1, 2, ..., k . Define:

N_i = the number of males in the i-th subset,

A_i = the actual observed mean earnings of the males in the i-th subset,

E_i = the estimated mean earnings based on regression and factorial analysis as outlined for the i-th subset. This estimate approximates actual mean earnings,

P_i = the potential mean earnings for the i-th subset as given in Table V. This is the maximum of the estimated mean earnings over 10 occupational classes.

The corresponding values for "total earnings" in Table VI were obtained by summing over subsets as follows:

$$\text{Total Actual Earnings} = \sum_{i=1}^{k} N_i A_i ,$$

$$\text{Total Estimated Earnings} = \sum_{i=1}^{k} N_i E_i ,$$

$$\text{Total Potential Earnings} = \sum_{i=1}^{k} N_i P_i .$$

The results in Table VII were obtained similarly, by finding the number of males having earnings below the potential median earnings for each subset and then summing over all subsets.

From these results it can be concluded that underemployment is minimum in the occupation class of managers, officials, and proprietors, except farm (i.e., "other managers" in Tables VI and VII). The worst group with respect to underemployment is farm laborers. Furthermore, the differential in underemployment between whites and nonwhites is greater in the South than in the North and West.

6. Application to North Carolina Survey Data. A socioeconomic survey was conducted in twelve selected counties of North Carolina in 1965. The information obtained for the sample individuals included age, race, sex, education, occupation and income during 1964. However, the samples were insufficient to make reliable estimates of mean and median incomes for the various subsets of interest. A preliminary analysis revealed that the reported incomes in 1964 in these counties were comparable with estimated earnings for similar individuals in the South for 1959. Hence, further calculations were made using the potential mean and median earnings for subsets computed for the South for 1959. The results are given in Tables VIII and IX.

It should be borne in mind that the underemployment problem in these counties in 1964 may be more severe than indicated by the figures in Tables VIII and IX, since the potential earnings for the South should have been higher in 1964 than in 1959. However, comparisons between counties are still possible. Thus, the worst counties are Avery and Yancey and the best counties are Halifax and Richmond.

7. Further Applications. In view of the high multiple correlations observed with the U. S. data, it appears that the number of variables considered may be adequate. Since information on these characteristics is also collected in the Current Population Surveys (CPS), underemployment for intercensal years can be evaluated. For example, underemployment in 1967 relative to 1959 (census year) can be measured by:

1. Converting estimated 1967 actual earnings to 1959 dollars,
2. Computing estimated 1967 potential earnings using 1959 potential mean earnings with 1967 weights for the appropriate subsets defined by age, race, education, region and occupation,
3. Converting the resulting estimated percent loss of output to an index by computing the ratio $(L_B/L_C) \times 100$, where L_B is the percent loss of output for the current year 1967 and L_C is the percent loss of output for the base year 1959. It is not essential that current actual earnings (or potential earnings) be converted to census year dollars; any other suitable base year can be chosen for this purpose. It is important, however, that the potential mean and median earnings for the age, race, education, region and occupation subsets be computed using data for the most recent year for which reliable estimates of these quantities are available. Periodic up-dating of the potential mean and median earnings for the subsets of interest is required to keep abreast with the effects of technological and social change, as reflected in the changing demand for specific occupational skills and in the removal of barriers to full employment. Sufficient data for up-dating potentials may only become available after each decennial census.

Although not carried out with the twelve North Carolina counties, a procedure similar to the above can be used for comparisons between communities, counties, state economic areas, or other areas of interest. Briefly, L_B is computed using census data for the specific areas to determine the base year weights (age, race, education, occupation subset sizes), but using regional potential mean earnings. L_C is computed, using current survey data to estimate total earnings and to estimate the current weights, with the same base year regional potential means. This quality is the same as computed for the twelve North Carolina counties. Comparisons between areas are made using the standardized index $100L_C/L_B$ for each area.

The suggested underemployment indices will reflect the characteristics of economic cycles as well as the specific structural-frictional aspects of various communities. To distinguish structural underemployment from cyclical underemployment, it is necessary to standardize for the cyclical sensitivity of area underemployment. An approach similar to that used by Miller (1968) on area unemployment could be used. The proposed indices may be useful for testing economic theories and hypotheses concerning occupational wage differentials and their changes over the business cycle. Particular reference is made to the alternate hypotheses of Reder (1955) and Oi (1962). The comparison of indices of underemployment and unemployment may shed additional light on these hypotheses.

8. Conclusions. No claims are made concerning the merits of the proposed technique for measuring the "absolute" level of underemployment. It seems clear that the proposed approach overestimates the "absolute" level of underemployment particularly in that it ignores (a) certain aspects of the labor market, (b) differences in individual potential earnings associated with innate abilities or with special skills obtained in training programs, and (c) personal occupational preferences. These deficiencies should not greatly deter the usefulness of the index for comparing populations in different areas or the same population in different time periods as a relative measure. The method has the advantage of flexibility in that it may be used to establish a consistent set of indices of underemployment over time and space for rural or urban populations, for occupation groups, and for various social classes defined by race, education, age and sex. Further, the data required to compute the index may be relatively inexpensive to obtain.

Census year data are used to estimate the required potential earnings standards and base year indices. Annual indices on a regional basis could be calculated using data collected by the Bureau of the Census Current Population Surveys. Current indices for populations confined to smaller areas such as counties or state economic areas may require surveys of sufficient size to estimate actual earnings and the weights (subset sizes) required by the method. The extent to which secondary sources might provide the needed data on current earnings and on the weights on a routine basis has not been explored.

9. Ackowledgement. The author wishes to thank Professors Dale M. Hoover and Robert M. Fearn of North Carolina State University and also Mr. Thomas Johnson of Research Triangle Institute for their comments and suggestions.

The author is indebted to the research division of the North Carolina Employment Security Commission for their cooperation and support in this work.

REFERENCES

[1] Benrud, Charles H., "Methodology for Estimating Unemployment in Rural Areas," Unpublished technical report No. 1 prepared for the State of North Carolina by Research Triangle Institute, Research Triangle Park, North Carolina, (1968).

[2] Bureau of the Census, Occupation by Earnings and Education, United Census of Population 1960, Final Report PC(2) - 7B, (1963).

[3] Libenstein, Harvey, "The Theory of Underemployment in Backward Economies," Journal of Political Economy, 65, (1957), 91-103.

[4] Miller, Edward, "The Separation of Structural from Cyclical Unemployment", Unpublished discussion paper No. 22, Program on Regional and Urban Economics, Harvard University, (1967).

[5] Nurske, R., Problems of Capital Formation in Underdeveloped Countries, Oxford University, England, (1953).

[6] Oi, Walter Y., "Labor as a quasi-fixed factor," Journal of Political Economy, 70, (1962), 538-555.

[7] Reder, Melvin W., "The theory of occupational wage differentials," American Economic Review, 45, (1955), 833-852.

[8] Report of the President's committee to appraise Employment and Unemployment Statistics, "Measuring Employment and Unemployment," U. S. Government Printing Office, Washington, D. C., (1962).

[9] United Nations Report, "Measures for the Economic Development of the Underdeveloped Countries," (1951).

TABLE I. List of characteristics with description

Factor A - Occupational class
1. Professional, Technical, and Kindred Workers
2. Farmers and Farm Managers
3. Managers, Officials and Proprietors, Except Farm
4. Clerical and Kindred Workers
5. Sales Workers
6. Craftsmen, Foremen and Kindred Workers
7. Operatives and Kindred Workers
8. Service Workers
9. Farm Laborers and Foremen
10. Laborers, Except Farm

Factor B - Sex
1. Male

Note: Females were not included in this analysis.

Factor C - Region (See Appendix D)
1. North and West United States
2. South United States

Factor D - Race
1. White
2. Non-white

Factor E - Age
1. 18-24 years
2. 25-34 years
3. 35-44 years
4. 45-54 years
5. 55-64 years

Factor F - Education

1.	Elementary School:	0-7 years
2.	Elementary School:	8 years
3.	High School:	1-3 years
4.	High School:	4 years
5.	College:	1-3 years
6.	College:	4 or more years

TABLE II. Preliminary analysis of variance of 1959 earnings data

Source	Degrees of Freedom	Mean: Sums of Squares	Mean: Mean Squares	Median: Sums of Squares	Median: Mean Squares
Main Effects:					
Occupation	9	1395.6	155.1	820.8	91.2
Race	1	57.3	57.3	58.6	58.6
Age	3	87.2	29.1	37.3	12.4
Education	5	738.7	147.7	364.7	72.9
Second-Order Interactions with Occupation	81	421.2	5.2	127.2	1.6
Error	357	133.7	0.4	30.7	0.1
Total	479	2873.0		1458.5	

Note: (1) This analysis does not include the South region and the age-group 18-24 years.
(2) The sums of squares and mean squares are in the units of 10^6.

TABLE III. Regression coefficients for the regressions of mean and median earnings for each occupational class

A. Mean Earnings

Source		Occupation Class 1	2	3	4	5	6	7	8	9	10
Mean		4423	3124	6379	4427	5721	4667	3992	3423	2023	3143
Effects:											
Race	(2)-(1)	-1746	- 752	-1396	- 372	-	- 720	- 479	- 456	- 146	- 270
Region	(2)-(1)	- 465	- 127	- 569	- 176	- 474	- 477	- 473	- 315	- 352	- 480
Age	(2)-(1)	- 821	143	-1159	- 114	- 252	67	45	168	122	158
	(3)-(1)	1083	606	687	715	1188	709	609	635	301	521
	(4)-(1)	1878	550	2158	873	1468	676	532	486	261	418
	(5)-(1)	1872	12	3072	748	1141	373	222	204	47	136
Education	(2)-(1)	-1726	-1184	-2393	- 555	-1352	- 788	- 317	- 468	- 175	- 104
	(3)-(1)	- 219	- 505	- 969	- 160	- 625	- 388	- 132	- 186	67	43
	(4)-(1)	339	- 24	85	144	156	10	146	156	415	380
	(5)-(1)	820	1343	1630	333	1088	294	330	147	222	131
	(6)-(1)	3297	1932	4730	1275	3000	2041	659	1146	-	-
Multiple Correlation		.9295	.9630	.9252	.9661	.9650	.9702	.9661	.9464	.9517	.9638

B. Median Earnings

Source		Occupation Class 1	2	3	4	5	6	7	8	9	10
Mean		4383	2270	5045	4292	4749	4449	3820	3255	1782	3028
Effects:											
Race	(2)-(1)	-1054	- 461	- 927	- 292	-	- 695	- 442	- 404	- 87	- 214
Region	(2)-(1)	- 427	- 232	- 391	- 179	- 403	- 490	- 528	- 340	- 376	- 564
Age	(2)-(1)	- 257	229	- 363	63	187	155	123	290	166	218
	(3)-(1)	1038	492	867	716	1099	723	604	665	318	522
	(4)-(1)	1222	300	1264	761	1005	620	501	473	301	447
	(5)-(1)	1141	- 99	1123	595	412	312	205	189	- 32	169
Education	(2)-(1)	-1119	- 748	-1488	- 393	- 908	- 690	- 233	- 417	- 148	- 57
	(3)-(1)	- 58	- 274	- 520	- 62	- 407	- 281	- 49	- 155	42	63
	(4)-(1)	511	60	154	203	273	104	220	179	332	373
	(5)-(1)	690	877	999	232	773	297	301	181	245	26
	(6)-(1)	1668	1155	2990	856	2031	1627	346	968	-	-
Multiple Correlation		.9622	.9759	.9485	.9723	.9732	.9643	.9604	.9383	.9348	.9670

TABLE IV. Potential Mean Earnings (in dollars) and optimum occupation class for males 18 to 64 years old in the experienced civilian labor force with earnings in 1959, by race, region, years of school completed, and age for the United States

(a) White--North and West U. S.

Education	Age Groups 18 - 24	25 - 34	35 - 44	45 - 54	55 - 64
Elementary					
0-7 years	3081(6)	4992(6)	5756(3)	7244(3)	8051(3)
8 years	3252(6)	5259(6)	6550(3)	8112(3)	8998(3)
High School					
1-3 years	3596(6)	6416(3)	8153(3)	9635(3)	10473(3)
4 years	3826(6)	7366(3)	9207(3)	10585(3)	11496(3)
College					
1-3 years	4385(3)	8668(3)	10589(3)	12174(3)	12907(3)
4 or more	7072(3)	11093(3)	13416(3)	15267(3)	15968(3)

(b) Nonwhite--North and West U. S.

Education	18 - 24	25 - 34	35 - 44	45 - 54	55 - 64
Elementary					
0-7 years	2327(7)	3785(5)	5019(5)	5390(5)	5276(3)
8 years	2478(7)	4562(5)	5812(5)	6258(5)	6215(3)
High School					
1-3 years	2628(7)	5074(5)	6454(5)	6819(5)	7611(3)
4 years	2858(5)	5615(5)	7099(5)	7569(3)	8611(3)
College					
1-3 years	3544(5)	6698(5)	8263(5)	9275(3)	10140(3)
4 or more	5263(5)	8156(5)	10285(3)	12218(3)	13050(3)

(c) White--South U. S.

Education	18 - 24	25 - 34	35 - 44	45 - 54	55 - 64
Elementary					
0-7 years	2302(6)	3955(6)	4555(3)	5960(3)	6736(3)
8 years	2639(6)	4387(6)	5514(3)	6993(3)	7848(3)
High School					
1-3 years	2955(6)	5426(3)	7091(3)	8488(3)	9296(3)
4 years	3306(6)	6498(3)	8266(3)	9561(3)	10440(3)
College					
1-3 years	3955(3)	7980(3)	9829(3)	11330(3)	12032(3)
4 or more	6613(3)	10377(3)	12627(3)	14394(3)	15064(3)

(d) Nonwhite--South U. S.

Education	18 - 24	25 - 34	35 - 44	45 - 54	55 - 64
Elementary					
0-7 years	1375(7)	2658(4)	3777(5)	4064(5)	3729(3)
8 years	1691(7)	3557(5)	4736(5)	5097(5)	4834(3)
High School					
1-3 years	1944(4)	4042(5)	5350(5)	5631(5)	6202(3)
4 years	2218(4)	4705(5)	6117(5)	6313(3)	7324(3)
College					
1-3 years	3073(5)	5969(5)	7462(5)	8200(3)	9033(3)
4 or more	4763(5)	7398(5)	9292(5)	11114(3)	11915(3)

NOTE: Number in the bracket indicates optimum occupation group.

TABLE V. Potential Median Earnings (in dollars) and optimum occupation class for males 18 to 64 years old in the experienced civilian labor force with earnings in 1959, by race, region, years of school completed, and age for the United States

(a) White--North and West U. S.

	Age Groups				
Education	18 - 24	25 - 34	35 - 44	45 - 54	55 - 64
Elementary					
0-7 years	3246(6)	5108(6)	5475(6)	5214(1)	5435(3)
8 years	3363(6)	5321(6)	5733(6)	6170(3)	6542(3)
High School					
1-3 years	3607(6)	5733(1)	6985(1)	7460(3)	7766(3)
4 years	3894(6)	6317(1)	7650(1)	8049(3)	8419(3)
College					
1-3 years	3453(6)	7021(3)	8356(3)	9252(3)	9434(3)
4 or more	5253(3)	8635(3)	10265(3)	11341(3)	11432(3)

(b) Nonwhite--North and West U. S.

Education	18 - 24	25 - 34	35 - 44	45 - 54	55 - 64
Elementary					
0-7 years	1997(7)	3909(5)	4980(5)	5339(5)	4954(5)
8 years	2151(7)	4728(5)	5844(5)	6284(5)	5961(5)
High School					
1-3 years	2435(5)	5177(5)	6376(5)	6737(5)	6869(3)
4 years	2933(5)	5656(5)	6938(5)	7225(5)	7557(3)
College					
1-3 years	3592(5)	6888(5)	8177(5)	8561(5)	8981(3)
4 or more	5390(5)	8122(5)	9705(5)	10783(3)	11151(3)

(c) White--South U. S.

Education	18 - 24	25 - 34	35 - 44	45 - 54	55 - 64
Elementary					
0-7 years	2388(6)	3959(6)	4335(6)	4105(3)	4405(3)
8 years	2652(6)	4320(6)	4740(6)	5299(3)	5660(3)
High School					
1-3 years	2894(6)	4775(1)	6035(1)	6586(3)	6881(3)
4 years	3356(6)	5535(1)	6876(1)	7352(3)	7710(3)
College					
1-3 years	3012(6)	6492(3)	7836(3)	8651(3)	8821(3)
4 or more	5201(3)	8293(3)	9932(3)	10927(3)	11006(3)

(d) Nonwhite--South U. S.

Education	18 - 24	25 - 34	35 - 44	45 - 54	55 - 64
Elementary					
0-7 years	915(4)	2694(5)	3774(5)	4052(5)	3656(5)
8 years	1353(4)	3661(5)	4786(5)	5145(5)	4811(5)
High School					
1-3 years	1656(5)	4107(5)	5316(5)	5595(5)	5716(3)
4 years	2330(5)	4763(5)	6053(5)	6259(5)	6579(3)
College					
1-3 years	3085(5)	6091(5)	7388(5)	7692(5)	8100(3)
4 or more	5070(5)	7511(5)	9103(5)	10100(3)	10457(3)

NOTE: Number in the bracket indicates optimum occupation group.

TABLE VI. Estimated 1959 loss of output as a percent of potential earnings for two regions of U. S.

A. North and West U. S.

Occupation	No. of Males (000's)	Total Earnings in Million Dollars			Loss of Output L in Percent
		Actual*	Estimated**	Potential	
1. Professional	3215	27.4	28.4	35.3	22
2. Farm Managers	1182	4.5	4.5	9.9	55
3. Other Managers	3042	30.0	30.2	30.3	1
4. Clerical	2157	10.8	10.8	17.7	40
5. Sales	1836	12.6	12.6	16.7	25
6. Craftsmen	6162	35.2	35.4	49.5	29
7. Operatives	6204	29.5	29.6	44.5	34
8. Service	1679	6.7	6.6	12.7	47
9. Farm Labor	451	1.0	1.0	2.7	63
10. Other Labor	1880	7.1	7.1	12.4	43
Race					
1. White	26310	159.0	160.5	223.3	29
2. Non-white	1498	5.8	5.7	8.4	31
Total	27808	164.8	166.2	231.7	29

B. South U. S.

Occupation	No. of Males (000's)	Total Earnings in Million Dollars			Loss of Output L in Percent	Estimates Based on Potential Earnings for North	
		Actual*	Estimated**	Potential		Potential	Loss
1. Professional	995	8.03	8.01	10.06	20	10.08	26
2. Farm Managers	666	1.76	1.86	4.18	58	4.93	64
3. Other Managers	1214	10.03	10.46	10.51	5	11.67	14
4. Clerical	679	3.13	3.18	4.87	36	5.47	43
5. Sales	688	3.92	4.03	5.38	27	6.00	35
6. Craftsmen	2285	10.42	10.38	14.54	28	16.88	38
7. Operatives	2416	8.59	8.62	12.79	33	15.25	44
8. Service	602	1.77	1.73	3.24	46	3.88	54
9. Farm Labor	409	0.52	0.49	1.69	70	2.13	76
10. Other Labor	971	2.32	2.32	4.28	46	5.33	57
Race							
1. White	9152	46.59	47.21	64.76	28	73.59	37
2. Non-white	1773	3.90	3.87	6.78	42	8.81	56
Total	10925	50.49	51.08	71.54	29	82.40	39

* Actual refers to 1959 Census results.

** Estimated refers to results obtained using factorial analysis and regression equations.

TABLE VII. Estimated number of 1959 males with earnings below the potential median earnings and the percentage underemployed, by region, in the United States

North and West U. S.

Occupation	Total No. of Males (000's)	No. of Males (000's) With Earnings Below: Estimated Median	Potential Median	Percent Under-employed U
1. Professional	3215	1700	2077	15
2. Farm Managers	1182	578	1033	37
3. Other Managers	3042	1563	1639	4
4. Clerical	2157	1086	1759	32
5. Sales	1836	917	1241	18
6. Craftsmen	6162	3154	4094	16
7. Operatives	6204	3200	4685	26
8. Service	1679	838	1490	39
9. Farm Labor	451	259	439	47
10. Other Labor	1880	954	1581	16
Race				
1. White	26310	13533	18797	21
2. Non-white	1498	716	1241	33
Total	27808	14249	20038	28

South U. S.

Occupation	Total No. of Males (000's)	No. of Males (000's) With Earnings Below: Estimated Median	Potential Median	Percent Under-employed U	Estimates Based on Potential Medians for North: Number (000's)	U
1. Professional	995	516	653	16	710	21
2. Farm Managers	666	348	602	40	623	43
3. Other Managers	1214	641	674	6	769	13
4. Clerical	679	355	517	26	586	36
5. Sales	688	351	477	19	530	27
6. Craftsmen	2285	1133	1473	14	1797	29
7. Operatives	2416	1247	1847	26	2166	40
8. Service	602	299	516	36	568	44
9. Farm Labor	409	191	395	46	407	49
10. Other Labor	971	488	814	34	914	44
Race						
1. White	9152	4727	6462	21	738	31
2. Non-white	1773	844	1506	35	169	45
Total	10925	5571	7968	23	907	33

TABLE VIII. Estimated and potential earnings for twelve North Carolina counties

County	No. of Males in the Sample	Earnings in Thousand Dollars			Percent Loss of Output L
		Actual* 1964	Estimated** Earnings	Potential** Earnings	
Bertie	130	467	462	755	38
Halifax	136	527	471	721	27
Hertford	140	497	535	826	40
Northampton	125	366	438	691	47
Macon	125	432	513	797	46
Richmond	122	512	502	724	30
Robeson	120	367	367	608	40
Scotland	112	367	394	614	40
Watauga	140	480	601	897	46
Avery	72	206	294	445	54
Mitchell	137	495	631	892	45
Yancey	101	301	380	616	52
Total	1460	5010	5012	8579	42

* Based on incomes reported in the survey rather than on earnings.

** Computed using 1959 mean earnings and potential mean earnings for the age, race, education and occupation subsets in the South.

TABLE IX. Estimated number of males having income (actual, 1964) below estimated and potential median earnings for twelve North Carolina counties

County	No. of Males in the Sample	The Number With Income Below		Percent Under-employed U
		Estimated Median Earnings*	Potential Median Earnings*	
Bertie	130	53	101	28
Halifax	136	54	87	14
Hertford	140	64	104	24
Northampton	125	67	105	34
Macon	125	74	99	29
Richmond	122	52	83	18
Robeson	120	58	95	29
Scotland	112	63	88	29
Watauga	140	84	110	29
Avery	72	55	62	36
Mitchell	137	84	108	29
Yancey	101	65	87	36
Total	1460	773	1129	27

* Computed using 1959 median earnings and potential median earnings for the age, race, education and occupation subsets in the South.

INTERNAL NET MIGRATION RESPONSE DIFFERENTIALS FOR THE UNITED STATES BY NON-LINEAR ITERATIVE LEAST SQUARES ESTIMATION PROCEDURE*

G. K. Kripalani, Western Michigan University

Introduction

This study is concerned with analyses of changes in internal net migration of small population groups stratified by age, sex and color for geographical sub-regions of the United States.

Several widely divergent motives may underlie the migration behavioral pattern of the people of an area. Better wages, or more generally, more favorable economic opportunities, present or potential, represent one major group of factors influencing migration decisions. Another major group of causes stems from socio-cultural environment of the areas of origin of migrants and their anticipated evaluation of corresponding socio-cultural situations in the areas of potential in-migration. Migration decisions are also affected by information, costs, existence of programs of assistance and kindred factors.

A theory of labor migration which regards relative wage ratio as the sole primary determinant of net migration is considered too simple and too inadequate to be useful in theoretical formulations or in empirical investigations. The general remarks in the preceding paragraphs suggest that neither relative wage ratio nor even some of the major relative economic opportunity factors may completely explain migration behavior. Moreover, for some of the major independent variables included in the model, adequately valid data series may not be available; besides, the nature of some particular variable may be such as to preclude its measurability or observability. For example, valid reasons are advanced that relative wage ratio, which by general consensus is regarded as a major economic variable in any model explaining net migration, is nonobservable because the ratio should relate to the marginal workers confronting migration choice and should not be the ratio of average wages.

The principal premise underlying this study is that net migration is a function of more than one major variable even when the analyses are based on data subdivided into small age-sex-color groups, that some of these variables are nonmeasurable or nonobservable and that valid data series for such variables do not exist for use in empirical investigations. It is in this basic premise of the nonobservable character of some of the major explanatory variables that the principal justification for the method of analyses used in this study lies. The procedure used does not permit testing of hypothesis but yields estimates of model parameters which can be interpreted in terms of the hypothesis underlying model specification.

In general terms, a theory of supply of net migrants may be formulated as follows: the number of net migrants from area A to the rest of the nation or to area A by the rest of the nation is a function of several variables, some of which are measurable and some of which are not. It will be assumed that nonmeasurable variables have a significant role to play in explaining variations in the supply of net migrants. Mathematically, we may write:

$$Y = f\ (Z_1, Z_2, \ldots, Z_n)$$

where Y represents the supply of net migrants and $Z_1, Z_2, \ldots, Z_n$ represent the independent variables such as relative wage ratio, other economic opportunity factors, phase of the business cycle and non-economic factors influencing internal migration.

Since some of the variables are not measurable, we may replace all variables, measurable and nonmeasurable, by one "omnibus" variable by means of the substitution $g(Z) = f\ (Z_1, Z_2, \ldots, Z_n)$. Hence we have:

$$Y = g\ (Z)$$

For example, if initially, we considered

$$Y = \alpha\ Z_1^{\beta_1} Z_2^{\beta_2} \ldots Z_n^{\beta_n}\ e$$

as the complete specification of the model, we would use the substitution

$$Z^\beta = Z_1^{\beta_1} Z_2^{\beta_2} \ldots Z_n^{\beta_n}\ e$$

and consider the model

$$Y = \alpha\ Z^\beta\ e$$

where Z is nonmeasurable "omnibus" variable, and e, the random term. The "omnibus" variable Z may be regarded as representing an index of relative opportunity which incorporates all economic and non-economic variables affecting net migration. The problem of estimating Z and the model parameters α and β has been handled by Johnston and Tolley (1968) by using a nonlinear iterative procedure, a slightly modified version of the NILES procedure of Wold (1965). In this study Johnston and Tolley formulation has been used. An essential property of this interative procedure is that the sequence of parameter estimates obtained at various iteration stages converges to the underlying parameter value not in an absolute sense but in a relative sense. Hence the character of Z, α and β estimates obtained at the final iteration stage is not cardinal but ordinal.

Net migration behaviour model of this study is based on the hypothesis that within a color-sex category the net migration rate of an age group is a function of nonobservable index of relative opportunity. Further, all age groups within a color-sex category face the same index at the same time. If suffix i refers to an age group within a color-sex category and suffix t refers to time, the model becomes

$$Y_{it} = \alpha_i Z_t^{\beta_i} \varepsilon_{it}$$

which after taking logarithms is transformed to linear form

$$Y_{it}' = \alpha_i' + \beta_i Z_t' + \varepsilon_{it}'$$

The dependent variable Y_{it} is the logarithm of net migration rate Y_{it} calculated as

$$Y_{it} = (E_{it} + M_{it})/E_{it}$$

where M_{it} = number of net migrants of age group i during time period t of the color-sex category in question, and E_{it} = the population exposed to risk of net migration of age group i during time period t of the color-sex category in question.

Y_{it} represents the 'survival' rate against net migration on the analogy of life table survival rates in the theory of single and multiple decrement tables.

Data Required and Their Sources

Statistical data required in this study related to the number of net migrants, M_{it} and the number exposed to risk of net migration, E_{it} used as the appropriate supply shifter. The available material regarding the estimated number of net migrants and the net migration rate for individual color-sex-age groups using census-survival rates was considered unsatisfactory. The problem of estimating the number of net migrants and the net migration rate was investigated independently of the existing procedures and formulas and an alternative approach devised. See Kripalani (1968).

It was recognized that the amount of computational work involved in preparing the basic statistics required in respect of all states for each of the four color-sex categories (white male, white female, nonwhite male and nonwhite female) for small age groups for six decades was very large. Practical considerations therefore suggested that one of the existing series of data would have to be used. It was decided to use the estimates prepared by Lee et al. (1957) in their momentous work, Population Redistribution and Economic Growth, United States, 1870-1950, for the five decades 1900-10 to 1940-50. For 1950-60, however, considerations of comparability warranted that corresponding estimates be prepared on the basis of formulas used by Lee. Lee's estimates of net migration (numbers and rates) are available for each of the four color-sex categories with five broad age groups 0-4, 5-14, 15-34, 35-54, and 55+ at the beginning of the decade (or 10-14, 15-24, 25-44, 45-64, and 65+ at the end of the decade).

The cross-sectional analyses for the 1950-60 decade were based on statistics published by the Economic Research Service of the U. S. Department of Agriculture (1965) in their Population-Migration Report giving net migration numbers and rates by age, sex and color separately for metropolitan and nonmetropolitan state economic areas. The published data were aggregated, wherever necessary, for the age groups 0-9, 10-14, 15-19, 20-24, 25-34, 35-44, 45-54, 55-64, and 65+ at the start of the decade for each of the four color-sex categories.

The iterative procedure was applied to Johnston and Tolley model

$$Y_{it}' = \alpha_i' + \beta_i Z_t' + e_{it}'$$

to obtain estimates of net migration response differentials β's and the non-observable omnibus variable Z's by age, sex and color for the 48 states of the Continental United States and the District of Columbia, (a) based on analyses of time series data for the six decades 1900-10 through 1950-60 and (b) based on analyses of cross section data for 1950-60 decade separately for metropolitan state economic areas (MSEA) and non-metropolitan state economic areas. Analyses were made separately for each of the four color-sex categories--white male, white female, nonwhite male, and nonwhite female. For time series analyses, each category was further subdivided into five age groups based on Lee's data, while for cross-section analyses, each category was subdivided into nine age groups based on USDA data.

Empirical Results of Time Series Analyses

In about 50 percent of the cases, 5 or less than 5 iterations were required to terminate the iterative process. In about 20 percent of the cases, more than 10 iterations were necessary. Total sum of squares explained by the regression equation at the final iteration state was 90 percent or higher in 64 percent of the cases; R^2 was below 50 percent in 17 percent of the cases.

Estimates of Omnibus Variable Z_t' over Time. The ordinal character of parameter estimates permitted the ranking of the index of relative opportunity. The rank analyses showed that there were significant differences in the ranks of the index of relative opportunity between the six decades spanned by the study. The test statistic (Friedman's X^2_r -See Siegel (1956), p. 166-) was generally highly significant for the United States as a whole, and for some important net in-migration and net out-migration regions. (Tables 1 and 2).

Main findings for net in-migration data (Table 1) were: For white male and female categories, Z_t' was highest in 1900-10 decade and touched its lowest value during the 1930-40 depression decade, a result which might have been expected. This is evidence of the fact that during the depression decade the index of relative opportunity which is above unity in net in-migration areas tends to move toward unity or there is a reduction in the strength of the pull forces exercised by an in-migration area on the potential migrants in the rest of the nation. The index progressively declined from 1900-10 to 1930-40 showing that the wider dispersion of economic activity during the World War I and post-World War I decades reduced the attractiveness of traditional net in-migration states for white males and white females. As compared to the depression decade, the relative attractiveness of net in-migration states improved significantly in the subsequent two decades but remained at levels substantially below that of pre-depression decades.

For the nonwhites, the increased economic activity of World War I and the following decade greatly increased the relative attractiveness of the net in-migration states, probably reflecting the fact that during these two decades, the growth of economic opportunities in traditionally nonwhite net out-migration areas did not decrease the incentive for nonwhites to move out as it did for the whites. Contrarily, the strength of the pull exercised by net in-migration states on potential nonwhite migrants increased during these decades. The depression decade considerably reduced the strength of pull forces on the nonwhites as it did for the whites. The index of relative opportunity for net in-migration states for nonwhites picked up again in the 1940-50 decade and the 1950-60 decade from the low levels of the depression decade, but the index was lower than the levels of 1910-20 and 1920-30. A possible reason which might explain this phenomenon could be that some of the main net in-migration states which were important for nonwhites in the early part of the present century might have ceased to grow at the relatively high rate as previously. Another possible reason could lie in the increasing nonwhite population base of these net in-migration states. Reduction in the attractiveness of net in-migration states for the nonwhites could also possibly arise from (a) a rise in the average age of the potential nonwhite out-migrants due to changes in population age composition and (b) increased skill requirements in jobs due to technological developments in the 1940-50 and 1950-60 decades as compared to 1910-20 and 1920-30 decades.

Main findings for net out-migration data (Table 2) were: the index of relative opportunity worsened continuously over the six decades except for the reversal witnessed during the depression decade for white males and white females. The index, which has value less than unity in the case of net out-migration states, had the highest rank in the first decade of the century and lowest rank in 1950-60 decade for white males. The highest rank for white females occurred in 1910-20 and the lowest rank in 1950-60. Similarly in the case of nonwhites the lowest rank was observed in the 1950-60 decade, thus showing that the index of relative opportunity tended to worsen for whites and nonwhites of both sexes over the period spanned by the study or the strength of push forces operating on potential migrants in net out-migration states was greater in the 1950-60 decade than in any previous decade of the present century.

The depression decade witnessed a considerable decrease in the strength of push forces operating in net out-migration states in the case of all four color-sex categories; except for white males, the strength of push forces was lowest in the depression decade for the other three categories.

From the analyses pertaining to the net in-migration and net out-migration data, there is adequate support for the hypothesis that during a period of general decline in aggregate demand, human resource adjustment process is greatly slowed down. When national unemployment rate increases, the number of both net in-migrants and net out-migrants falls; hence the dependent variable falls in the case of net in-migration and rises in the case of net out-migration. This is reflected in the reduced attractiveness of net in-migration states and increased attractiveness of net out-migration states for potential migrants.

Estimates of Age Group Response Coefficient, β_i (Tables 3 and 4). Comparative inferences about how different age groups respond to changes in the index of relative opportunity could be made on the basis of η's, $\eta_{ij} = \beta_i/\beta_j$ because the iterative procedure yielded estimates of response coefficients (elasticities) which were ordinal in character. Since the overall analyses were in terms of the structure of the communities of origin of net out-migrants and of destination of net in-migrants, it was expected that behavioral response coefficients would be significantly different in the case of a state for which data permitted both types of analyses. The base population in the case of net out-migrants is the population of the area from which net out-migration is taking place; net in-migrants, on the other hand, are related to the population of the area to which they go and not to the population base which forms the source of their supply. The empirical results showed that this expectation was reasonable.

Response coefficients differed significantly from age group to age group. The test statistic was highly significant at less than .1 percent level in all the four color-sex categories in both net in-migration and net out-migration data analyses. The substantive findings were:

Considering net in-migration for white male and female categories (Table 3), the response coefficient was highest for the age group 15-34 (at the start of the decade) followed by age group 5-14 and then by age group 0-4. This high level of response by age group 5-14 and 0-4 may be expected because these groups are generally not independent movers but are linked with their parents' migration decisions. The fourth in rank was age group 35-54 followed by 55+. In the case of nonwhite male and female categories (Table 3), the highest response coefficient was observed for the age group 5-14, perhaps an interesting evidence of the fact that nonwhite male and female migrants are younger than the white migrants. Led by age group 5-14, the other age groups in decreasing order of their response coefficients were 15-34, 0-4, 35-54 and 55+.

Considering net out-migration in these analyses (Table 4), the order pattern was the same for all the four color-sex categories, age group 15-34 having the highest coefficient, followed in order by 5-14, 0-4, 35-54 and 55+. In contrast to net in-migration analyses, the nonwhite male and female categories consistently showed rank 1 for the age group 15-34 in all the states covered. A possible explanation for the observed feature in the case of nonwhites, viz. highest β for age group 5-14 for net in-migration analyses but not for net out-migration, may be in fact that nonwhites are coming from the scattered areas in the South but are going to few northern states, giving rise to different population bases for different age groups.

Empirical Results of Cross Sectional Analyses

The substantive results of cross sectional analyses based on metropolitan state economic areas (MSEA), which are mainly urban net in-migration areas, and nonmetropolitan state economic areas (NSEA), which are mainly rural net out-migration areas, separately for the 1950-60 decade are briefly discussed below.

Main Results Based on MSEA Net In-migration Analyses (Part (a), Table 5)

Estimates of age group response coefficient, β_i. There was clear evidence that significant differences existed between the ranks of response coefficients for various age groups. The null hypothesis was rejected at .1 percent level in all the four color-sex categories.

The cross sectional analyses were based on smaller age group subdivisions, 9 in number as against 5 age groups in the time series analyses. The highest response coefficient in the case of whites was among age group 5 (ages 25-34 at the start of the decade) followed by age group 6 (35-44); in the case of nonwhites, however, the highest two ranks were shared by age groups 4 and 5 (20-24 and 25-34), a result indicative of the fact that among nonwhites the highest response to net migration opportunity factors occurs among relatively young age groups as compared to whites. In all the four categories, the lowest three ranks, 7, 8, and 9 for the response coefficients occurred among the youngest age group 1 (0-4) and the oldest two age groups 8 and 9 (55-64 and (65+).

Σ-Curve

Part (a) of Table 5 gives the ranks of β_i for the United States based on net in-migration analyses for MSEA. A Lorentz type curve, designated Σ-curve may be constructed, the curve being drawn by connecting points whose coordinates are

$$\left(i, \sum_{j=1}^{i} r_j\right)$$

where r_j is the rank for age group j. The Σ-curve corresponding to β_i is given in Figure 1.

Σ-curve may be used as a basis for a broad comparison between two categories. When Σ_A (Σ-curve for category A) lies below Σ_B for values of i up to, say, i = 6, one may reasonably infer that high ranks (low numerical values) occur more frequently among relatively young age groups in category A than in category B.

The highest response coefficient in the case of whites was among age group 5 followed by age group 6; in the case of nonwhites, however, the highest two ranks were shared by age groups 4 and 5. This result might be taken as indicative of the fact that within the two categories of nonwhites the highest response occurred among relatively young age groups as compared to the white categories. In all the four categories, the lowest three ranks 7, 8 and 9 for the response coefficients occurred for the youngest age group 1 and the oldest two age groups 8 and 9.

Table 6 shows the significance levels of X_r^2 for different regions. In the case of net in-migration data analyses, there is no evidence for rejecting the null hypothesis at 20 percent level in (a) East South Central and Pacific regions for white males, (b) West South Central, Mountain and Pacific regions for white females and (c) Middle Atlantic region for nonwhite females. For these regions, therefore, there was no evidence for the hypothesis that ranked elasticity of response with respect to omnibus variable differed significantly among the various age groups. On the other hand, the null hypothesis was rejected at the 5 percent level in (a) West South Central region for white males, (b) South Atlantic for white females, (c) Middle Atlantic, East North Central and West South Central for nonwhite males, and (d) East North Central for nonwhite females. Thus ranked response elasticities among various age groups differed significantly and displayed a definite ranking pattern among white male net in-migrants moving into the metropolitan state economic areas in the West South Central region. Similar observations might be made for regions cited above for relevant categories. Existence of a definite ranking pattern criterion, X_r^2 significant at 1 percent level, classification e in Table 6, might be taken as indicative of the fact that the region in question was more selective of some age groups rather than others of the category. Classification a or b, on the other hand, reflects that no strong evidence of the region being definitely more selective of some age groups as compared with others. For example, Middle Atlantic was definitely age selective of nonwhite males but not so of non-white females. East North Central was age selective for both sexes among the nonwhites. Quite significant were the results for the Pacific region, whose classification "a" (X_r^2 not significant at 20 percent level) showed that the region did not have a definite selection preference for some age groups relative to others among white male and white female net in-migrants moving into the metropolitan state economic areas of the region.

Rank analyses of β_i were of particular significance for Florida. The ranks for response coefficients for the oldest two age groups i = 8 and i = 9 were 3 and 1, respectively, for white males and 1 and 2, respectively, for white females, reflecting the highest net in-migration response by white pensioners of both sexes. The corresponding ranks were 6 and 8 for nonwhite male and 8 and 7 for nonwhite female categories.

Main Results Based on NSEA Net Out-migration Analyses (Part (b), Table 5)

Estimates of age group response coefficient, β_i. There was clear evidence of the existence of significant differences between ranks of response coefficients for various age groups. The null hypothesis of no significant differences between ranks was rejected at .1 percent level in all the four color-sex categories. Generally the highest response coefficient occurred for age group 20-24 followed by age group 25-34. The highest three ranks generally occurred among the three age groups 15-19, 20-24 and 25-34. The lowest three ranks were found to occur among the youngest age group 0-4 and the oldest two age groups 55-64 and 65+.

Part (b) of Table 5 gives the ranks of β_i for the United States based on net out-migration analyses of NSEA. Corresponding Σ-curve is given in Figure 2.

Table 7 gives the significance levels of X_r^2 in the rank analyses of β_i for different regions. Based on net out-migration data, the main results were:

(1) White male--the null hypothesis could be rejected at the 5 percent level in the East North Central, West North Central, South Atlantic, East South Central and West South Central regions. At the 10 percent level the null hypothesis could be rejected in all the regions except the Mountain and Pacific regions, both of which are regions of large net in-migration for white males.

(2) White Female--the null hypothesis could be rejected at the 5 percent level in West North Central, South Atlantic, East South Central, West South Central and Mountain regions. At the 10 percent level it could be rejected in all regions but two, these being New England and Middle Atlantic.

(3) Nonwhite male--in all the three main regions of net out-migration, viz. South Atlantic, East South Central and West South Central, the null hypothesis could be rejected at the .1 percent level.

(4) Nonwhite female--as in the case of non-white males, the null hypothesis could be rejected at the .1 percent level in all the three main regions of net out-migration.

A comparison of time series and cross section results yielded some interesting conclusion which might be expected. The highest response coefficient occurred among young persons and the lowest response coefficient among the old age groups. Further, the study of net in-migration data of MSEA which are predominantly urban areas and of net out-migration data of NSEA which are predominantly rural communities did not show significant pattern differences between response coefficients for these two types of areas, viz. rural and urban areas.

LIST OF REFERENCES

Agricultural Policy Institute. 1961. The farmer and migration in the United States. API Series No. 3, School of Agriculture, North Carolina State College, Raleigh.

Bailey, W. G. and Haycocks, H. W. 1946. Some Theoretical Aspects of Multiple Decrement Tables. T. and A. Constable, Ltd., Edinburgh, Scotland.

Bishop, C. E. 1961. Economic aspects of changes in farm labor force, pp. 36-49. In Labor Mobility and Population in Agriculture. Iowa State University Press, Ames.

Bunting, R. L. 1962. Labor mobility and wage improvement, pp. 208-219. Conference on Human Resources in the Urban Economy. Resources for the Future, Inc., Washington, D. C.

Diehl, W. D. 1964. Farm-nonfarm migration in the Southeast: A costs-returns analysis. Unpublished Ph.D. thesis, Department of Agricultural Economics, North Carolina State of the University of North Carolina at Raleigh. University microfilms, Ann Arbor, Michigan.

Fisher, R. A. and Yates, F. 1948. Statistical Tables for Biological, Agricultural and Medical Research. Oliver and Boyd, Ltd., Edinburgh, Scotland.

Jansen, E. F., Jr. 1964. Employment-to-population ratios. Unpublished Master's thesis, Department of Agricultural Economics, North Carolina State of the University of North Carolina at Raleigh.

Jansen, E. F., Jr. 1966. Employment participation behavioral relationships. Unpublished Ph.D. thesis, Department of Economics, North Carolina State University at Raleigh, University Microfilms, Ann Arbor, Michigan.

Jerome, H. 1926. Migration and Business Cycles. National Bureau of Economic Research, Inc., New York.

Johnston, W. E. and Tolley, G. S. 1968. The supply of farm operators. Econometrica, Vol. 36, No. 2 (April, 1968) pp. 365-382.

Kripalani, G. K. 1968. A statistical approach to the problem of estimating the number of internal net migrants and the internal net migration rates by census survival rate method. Presented as a contributed paper at the Social Statistics Section of the American Statistical Association, Pittsburgh, PA, August 20-23, 1968.

Lee, E. S., Miller, A. R., Brainerd, C. P., and Easterlin, R. A. 1957. Methodological considerations and reference tables, Volume I. Population Redistribution and Economic Growth, United States, 1870-1950. American Philosophical Society, Philadelphia.

Schultz, T. W. 1961. A policy to redistribute losses from economic progress. J. Farm Econ. 43 (3): 554-565.

Segal, M. 1962. Occupational wage differentials in major cities during the 1950's, pp. 195-207. Conference on Human Resources in the Urban Economy. Resources for the Future, Inc., Washington, D. C.

Siegel, S. 1956. Nonparametric Statistics for the Behavioral Sciences. McGraw-Hill Book Company, Inc., New York.

Sjaastad, L. A. 1961a. Income and migration in the United States. Unpublished Ph.D. thesis, Department of Economics, University of Chicago. University Microfilms, Ann Arbor, Michigan.

Sjaastad, L. A. 1961b. Occupational structure and migration patterns, pp. 8-27. In Labor Mobility and Population in Agriculture. Iowa State University Press, Ames.

U. S. Department of Agriculture. 1965. States, counties, economic areas and metropolitan areas, Volume I. Population-Migration Report: Net Migration of the Population, 1950-60, by Age, Sex and Color. U. S. Government Printing Office, Washington, D. C.

*The results reported in this paper are a part of author's work on a National Science Foundation Project "Area-Population Adjustments in relation to Economic Activity" at the North Carolina State University during 1963-65. The author gratefully acknowledges the debt to G. S. Tolley for his guidance at various critical stages in the study, also to C. E. Bishop, C. H. Hamilton and F. E. McVay for their valuable comments. Any errors or deficiencies that may yet remain are exclusively the responsibility of the author.

Table 1. Rank Analyses of $\hat{Z}_t$ for the United States and Regions based on Net Immigration Analyses of Time Series Data - Six Decades 1900-10 through 1950-60

Region	Number of states covered	Sum of ranks of $\hat{Z}'_t$ 1900-10	1910-20	1920-30	1930-40	1940-50	1950-60	Significance Level of Friedman's X_r^2 [a]
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
				(i) White Male				
United States	19	46	68	72	81	65	67	.10
New England	2	2	6	9	12	7	6	.20
Middle Atlantic	1	1	3	2	6	5	4	--
East North Central	2	9	2	6	11	7	7	n. s.
South Atlantic	5	25	19	19	13	14	15	n. s.
West South Central	3	3	15	10	11	13	11	.20
Mountain	3	3	11	16	15	11	7	.05
Pacific	3	3	12	10	13	8	17	.10
				(ii) White Female				
United States	20	42	58	65	93	80	82	.001
New England	4	5	10	17	19	16	17	.10
Middle Atlantic	2	2	6	4	10	10	10	.20
East North Central	3	10	5	6	16	14	12	.20
South Atlantic	3	14	9	13	9	10	8	n. s.
West South Central	2	3	10	3	9	9	8	n. s.
Mountain	3	4	9	13	14	13	10	n. s.
Pacific	3	4	9	9	16	8	17	.05
				(iii) Non-White Male				
United States	19	79	43	57	93	63	64	.001
New England	3	11	5	15	16	8	8	.20
Middle Atlantic	3	11	5	4	17	11	15	.05
East North Central	5	24	6	11	29	13	22	.001
West North Central	3	13	6	11	16	11	6	.20
South Atlantic	3	14	15	10	7	10	7	n. s.
West South Central	2	6	6	6	8	10	6	n. s.
				(iv) Non-White Female				
United States	19	81	47	51	82	66	72	.01
New England	3	13	9	15	12	8	6	n. s.
Middle Atlantic	3	12	7	3	16	11	14	.10
East North Central	5	26	8	9	28	13	21	.001
West North Central	3	11	5	7	11	14	15	n. s.
South Atlantic	3	14	14	10	8	8	9	n. s.
West South Central	2	5	4	7	7	12	7	n. s.

[a] n. s. - not significant at 20 percent level.

Table 2. Rank Analyses of $\hat{Z}'_t$ for the United States and Regions based on Net Outmigration Analyses of Time Series Data - Six Decades 1900-10 through 1950-60

Region	Number of states covered	Sum of ranks of $\hat{Z}'_t$ 1900-10	1910-20	1920-30	1930-40	1940-50	1950-60	Significance Level of Friedman's X^2_r [a]
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
			(i) White Male					
United States	22	58	67	83	71	84	99	.02
New England	4	6	11	16	15	17	19	.20
Middle Atlantic	2	2	8	7	5	10	10	n. s.
East North Central	2	7	3	5	7	10	10	n. s.
West North Central	3	9	8	9	9	15	13	n. s.
South Atlantic	3	15	11	12	12	6	7	n. s.
East South Central	4	10	13	13	8	16	24	.05
West South Central	2	5	8	9	6	6	8	n. s.
Mountain	2	4	5	12	9	4	8	.20
			(ii) White Female					
United States	20	59	51	78	53	89	90	.001
New England	3	5	8	16	7	13	14	.10
Middle Atlantic	1	1	2	3	4	5	6	--
East North Central	3	12	6	11	7	13	14	n. s.
West North Central	4	14	9	14	14	19	14	n. s.
South Atlantic	4	17	11	17	10	17	12	n. s.
East South Central	4	9	13	13	8	17	24	.05
West South Central	1	1	2	4	3	5	6	--
			(iii) Non-White Male					
United States	10	31	29	37	17	47	49	.001
South Atlantic	5	16	11	24	10	19	25	.05
East South Central	3	7	12	8	5	16	15	.10
West South Central	2	8	6	5	2	12	9	.20
			(iv) Non-White Female					
United States	10	26	25	41	26	46	46	.01
South Atlantic	5	13	11	23	17	18	23	n. s.
East South Central	3	6	9	12	6	16	14	.20
West South Central	2	7	5	6	3	12	9	n. s.

[a] n. s. - not significant.

Table 3. Rank Analyses of β_i for the United States and Regions based on Net In-migration Analyses of Time Series Data-Six Decades 1900-10 through 1950-60

Region	Number of States Covered	Sum of Ranks of β_i: β_1	β_2	β_3	β_4	β_5	Significance level of Friedman's X^2_r [a]
1	2	3	4	5	6	7	8
(i) White Male							
United States	16	45	39	26	57	73	.001
East North Central	4	10	10	9	15	16	n. s.
South Atlantic	2	3	7	3	7	10	.20
West South Central	2	5	7	4	5	9	n. s.
Mountain	5	18	9	7	18	23	.01
Pacific	3	9	6.	3	12	15	.02
(ii) White Female							
United States	23	72	56	34	82	101	.001
New England	2	7	2	4	9	8	.20
Middle Atlantic	1	3	1	2	4	5	--
East North Central	2	6	4	2	8	10	.10
West North Central	1	3	4	1	5	2	--
South Atlantic	4	13	11	4	12	20	.02
West South Central	3	7	6	9	9	14	n. s.
Mountain	7	25	20	7	26	27	.01
Pacific	3	8	8	5	9	15	.20
(iii) Non-White Male							
United States	18	46	30	34	75	85	.001
New England	3	8	4	6	13	14	.05
Middle Atlantic	3	8	5	5	14	13	.05
East North Central	5	13	7	10	20	25	.01
West North Central	3	6	5	9	11	14	n. s.
South Atlantic	3	8	7	3	13	14	.05
West South Central	1	3	2	1	4	5	--
(iv) Non-White Female							
United States	18	45	36	44	66	79	.001
New England	2	3	7	6	6	8	n. s.
Middle Atlantic	3	7	4	7	12	15	.05
East North Central	5	10	7	13	20	25	.01
West North Central	3	10	8	9	10	8	n. s.
South Atlantic	4	12	9	7	14	18	.20
West South Central	1	3	1	2	4	5	--

[a] n. s. - not significant at 20 percent level.

Table 4. Rank Analyses of β_i for the United States and Regions based on Net Out-migration Analyses of Time Series Data - Six Decades 1900-10 through 1950-60

Region	Number of State Covered	Sum of Ranks of β_i β_1	β_2	β_3	β_4	β_5	Signigicance level of Friedman's χ_r^2 [a]
1	2	3	4	5	6	7	8
		(i) White Male					
United States	26	84	65	44	96	101	.001
New England	4	12	12	4	16	16	.05
Middle Atlantic	1	3	1	2	4	5	--
East North Central	2	7	6	5	6	6	n. s.
West North Central	7	23	15	13	31	23	.02
South Atlantic	3	14	6	3	11	11	.05
East South Central	4	9	9	6	16	20	.01
West South Central	1	3	1	2	5	4	--
Mountain	4	13	15	9	7	16	n. s.
		(ii) White Female					
United States	26	72	55	54	96	113	.001
New England	3	9	7	5	9	15	.20
Middle Atlantic	1	4	2	1	5	3	--
East North Central	2	3	5	4	9	9	.20
West North Central	7	20	13	13	28	31	.01
South Atlantic	4	12	9	7	15	17	.20
East South Central	4	9	6	9	17	19	.02
West South Central	1	3	1	2	4	5	--
Mountain	4	12	12	13	9	14	n. s.
		(iii) Non-White Male					
United States	10	33	22	10	38	47	.001
South Atlantic	5	16	10	5	19	25	.001
East South Central	4	14	10	4	15	17	.05
West South Central	1	3	2	1	4	5	--
		(iv) Non-White Female					
United States	9	30	19	10	34	42	.001
South Atlantic	5	16	9	6	19	25	.001
East South Central	3	11	8	3	11	12	.20
West South Central	1	3	2	1	4	5	--

[a] n. s. - not significant at 20 percent level.

Table 5. Rank Analyses of β_i for the United States based on (a) Metropolitan State Economic Areas Net Inmigration Data and (b) Nonmetropolitan State Economic Areas Net Outmigration Data for 1950-60 Decade.

Category	Number of States Covered	Rank of β_i									Significance Level of Friedman's X^2_r
		i=1	i=2	i=3	i=4	i=5	i=6	i=7	i=8	i=9	
		(a) MSEA - Net Inmigration Analyses									
White Male	18	7	4	6	3	1	2	5	8	9	.001
Σ^a	-	7	11	17	20	21	23	28	36	45	--
White Female	17	7	3.5	5.5	3.5	1	2	5.5	8	9	.001
Σ	-	7	10.5	16	19.5	20.5	22.5	28	36	45	--
Nonwhite Male	13	7	4	3	1	2	5	6	8	9	.001
Σ	-	7	11	14	15	17	22	28	36	45	--
Nonwhite Female	13	7	4	3	2	1	5	6	8	9	.001
Σ	-	7	11	14	16	17	22	28	36	45	--
		(b) NSEA - Net Outmigration Analyses									
White Male	31	7	4	3	1	2	5	6	8	9	.001
Σ	-	7	11	14	15	17	22	28	36	45	--
White Female	30	7	4	2	1	3	5	6	8	9	.001
Σ	-	7	11	13	14	17	22	28	36	45	--
Nonwhite Male	16	8	5	4	2	1	3	6	7	9	.001
Σ	-	8	13	17	19	20	23	29	36	45	--
Nonwhite Female	17	7	5	3	1	2	4	6	8	9	.001
Σ	-	7	12	15	16	18	22	28	36	45	--

Σ^a - denotes the progressive summation of ranks up to and including i.

Table 6. Significance levels of Friedman's X_r^2 in the rank analyses of β_i based on MSEA[a]. Net In-migration data for 1950-60 decade.

Region	White male	White female	Nonwhite male	Nonwhite female
New England				
Middle Atlantic			e	a
East North Central	b	c	e	e
West North Central				
South Atlantic	c	d		
East South Central	a			
West South Central	e	a	b	c
Mountain	c	a		
Pacific	a	a		
United States total	e	e	e	e

Table 7. Significance levels of Friedman's X_r^2 in the rank analyses of β_i based on NSEA[a]. Net In-migration data for 1950-60 decade.

Region	White male	White female	Nonwhite male	Nonwhite female
New England	c	a		
Middle Atlantic	c	a		
East North Central	d	c		
West North Central	e	e	b	a
South Atlantic	e	e	e	e
East South Central	d	e	e	e
West South Central	e	e	e	e
Mountain	a	d		
Pacific	a			
United States total	e	e	e	e

[a] a - not significant at 20 percent level; b - significant at 20 percent level; c - significant at 10 percent level; d - significant at 5 percent level e - significant at.1 percent level.

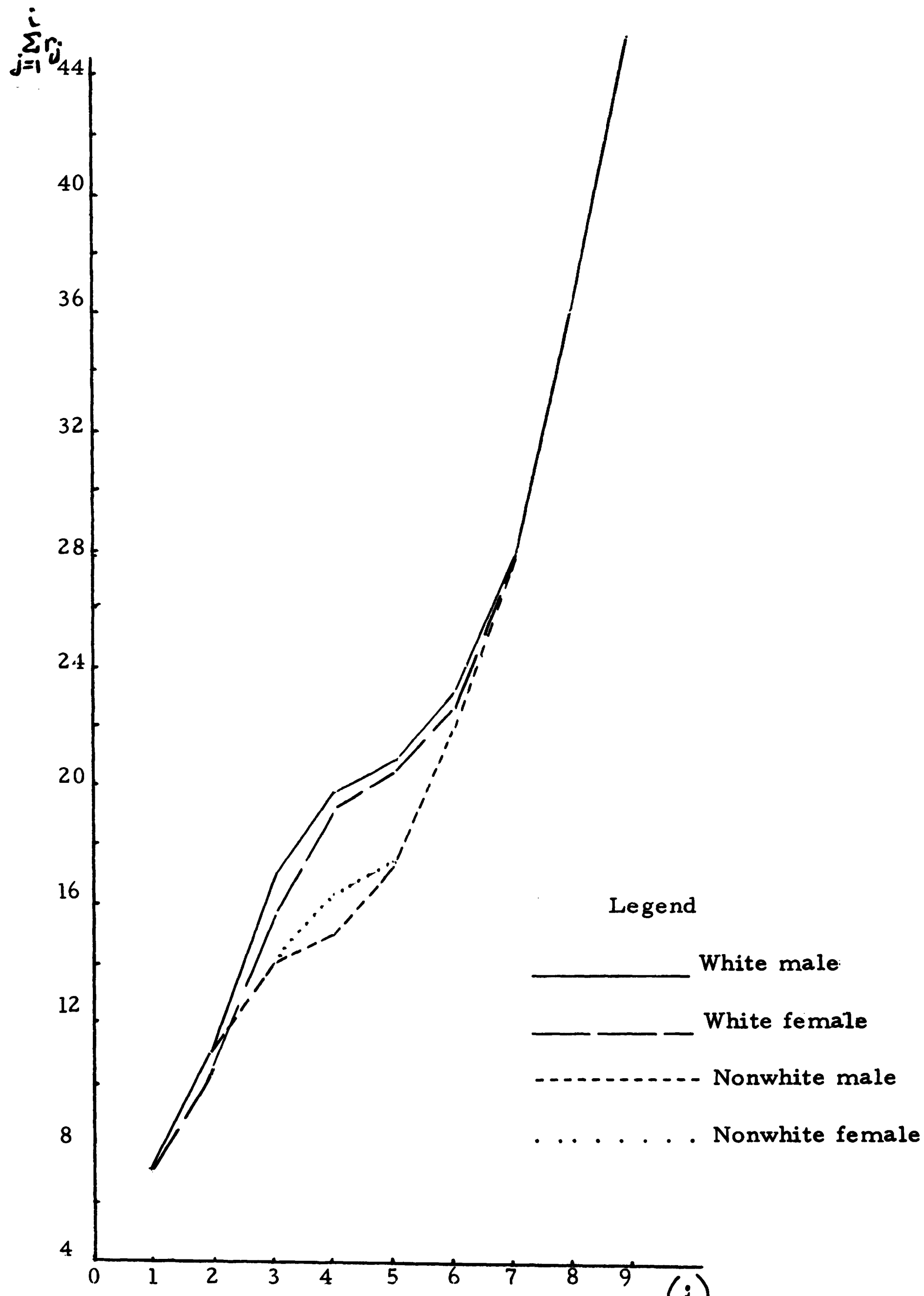


Figure 1.
$\sum$ curves for ranks of β_i' for different color-sex categories based on MSEA net in-migration data for 1950-60

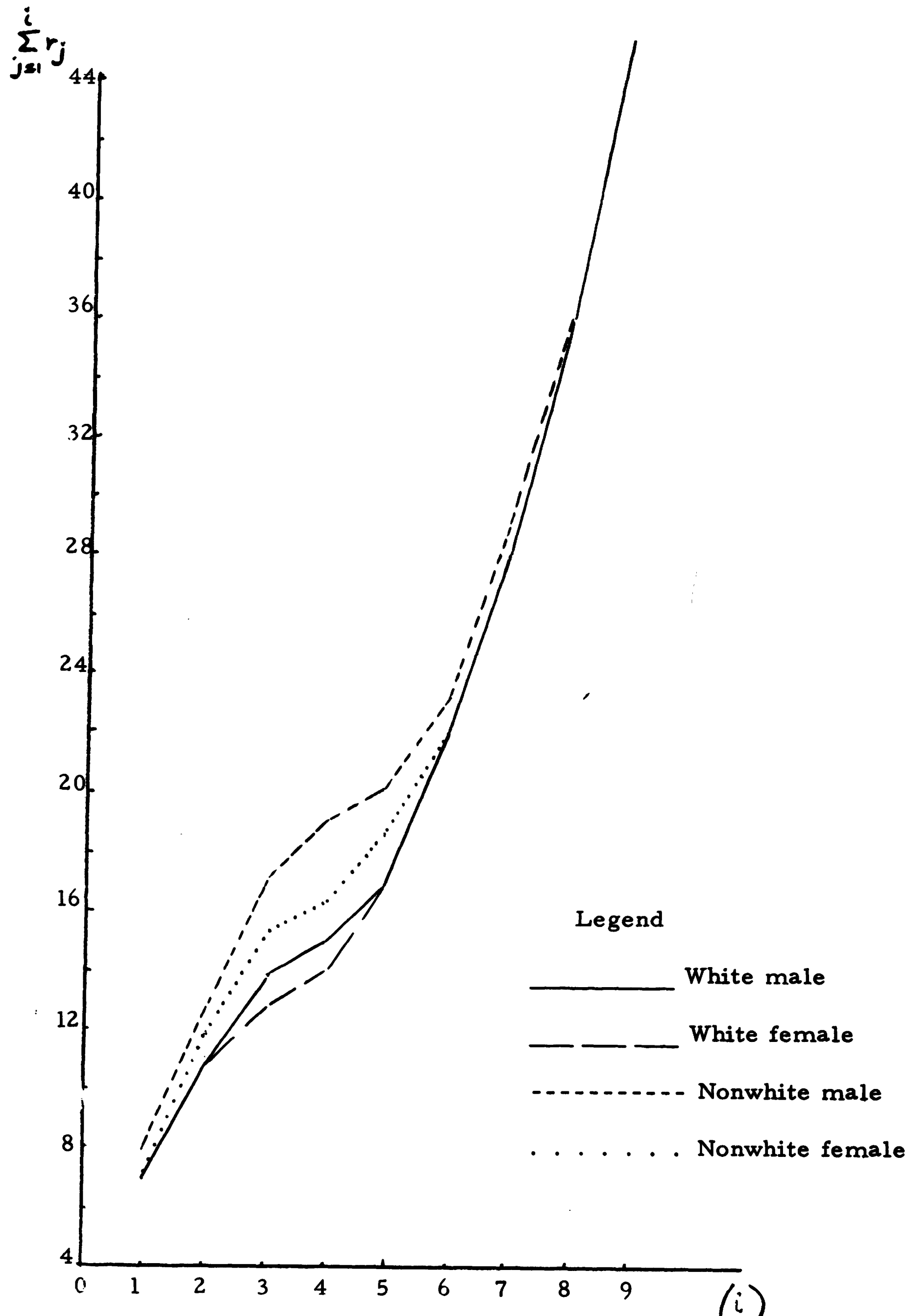


Figure 2.
$\sum$ - curves for ranks of β_i for different color-sex categories based on NSEA net out-migration data for 1950-60

USE OF SOCIAL SECURITY'S CONTINUOUS WORK HISTORY SAMPLE FOR MEASUREMENT OF NET MIGRATION BY GEOGRAPHIC AREA*

Meyer Zitter and Elisabeth S. Nagy
U. S. Bureau of the Census

This paper considers the potential of a newly developed resource, the Social Security Administration's Continuous Work History Sample, for use in derivation of estimates of net resident migration for geographic areas such as regions, States, and Standard Metropolitan Statistical Areas. The nature of the Work History File, its coverage, its limitations and the procedural problems associated with its use for measurement of resident migration are discussed. Estimates of net migration derived from the Work History File for regions, States, and SMSA's by race for 1960-65 are compared with closely corresponding estimates from other sources, such as the 1960 Census, postcensal surveys, and independent population estimates prepared at the Bureau of the Census.

Nature and coverage of the Work History File.--In any study involving a secondary use of administrative statistics one needs to consider most carefully applicability of the data to the problem at hand and the amount of manipulation necessary to make the data "fit the need." In the present case, a File containing incomplete employment data is examined for its applicability to the problem of measuring resident migration. As limiting as such a File may appear initially, it nevertheless constitutes an important source of information now regularly available to students of population movements.

The Continuous Work History Sample (CWHS) is a one-percent sample of all persons who have a Social Security account number and have worked in covered employment. The characteristics of persons in the Sample--age, sex, and race--come from the individual's application for a Social Security account number (Form SS-5). Information on earnings and employment is obtained from the quarterly earnings reports (annually in the case of farm workers) filed by each employer for his employees in covered employment (Forms IRS-941, IRS-942, IRS-943, and OAR-53). A separate sample file is maintained for self-employed persons. (The present study excludes the latter.)

The sample is selected on the basis of specified digits in the last four places of the nine-digit Social Security account number. Once an account number falls in the sample, it will reappear each year that the person works in covered employment. It is thus possible to make year-to-year comparisons for the same individual and to keep track of changes in his place of employment, coded by address of employer to State and SMSA.

In 1966 the workers in OASDHI programs comprised 88 percent of total civilian employment. The great majority of workers presently excluded from coverage fall into three categories: Federal civilian employees and some State and local government employees; household workers and farm workers who do not work long enough or earn enough to meet minimum requirements; and very low income self-employed persons.

Use of the File on the national level.--An overall summary of how well the Work History File reflects interstate migration of employed persons is provided in Table 1. In it, national summaries of the annual number of the Work History File's interstate migrants (reflecting change of place of employment and not necessarily change of residence) are compared with interstate migrants obtained from the Census Bureau's Current Population Survey (reflecting change of residence). 1/ Table 1 will be used to illustrate the basic problems of working with the File on the national level, although the description of the limitations and meaning of File data applies to all geographic levels and will be amplified when areas below the national level are discussed.

Length of migration period.--As Table 1 indicates, data from the Work History File were available both on a "calendar-year" and "first quarter-year" basis. The "calendar-year" tabulations include everybody who worked at any time during the year in covered employment. Geographic areas are assigned on the basis of the employer with whom the employee had maximum quarterly earnings. The "first quarter-year" figures cover only wage earners who worked in

*The programming and processing of the Work History Sample data tapes which provided the basic migration matrix for States and Standard Metropolitan Statistical Areas were carried out by David Hirschberg, Regional Economics Division, Office of Business Economics. Jerome M. Glynn, Population Division, Bureau of the Census, was responsible for further computer processing for adaptability to population estimation methodology. The assistance of Mildred Stanback in the statistical processing is gratefully acknowledged.

the first calendar quarter. (The self-employed are excluded in both instances. For the purpose of measuring migration, comparison should be made at identical points of the year in order to obtain change over a 12-month period. From this standpoint, first-quarter data are more appropriate than calendar-year statistics. Full calendar-year data for 1961 and 1962 are nonetheless used in this study in obtaining migration rates for the 1960-65 period, as first-quarter data for those years had not been reconstructed at the time of this writing. The inclusion in Table 1 of calendar-year statistics for 1961-64 and first-quarter statistics for 1963-65 serve to demonstrate the extent of the differences between the two time series. It is apparent, for example, that calendar-year data provide a substantially broader base to work from, since the total number of workers based on calendar-year data is 15-20 percent greater than the number derived from first-quarter data. The advantage which might be gained by using the larger number of sample cases is cancelled out, however, by greater uncertainty regarding the length of the migration period. Because of the overlap of the two series in 1963 and 1964, it is also possible to demonstrate rate differences which result from the use of different bases.

Area of coverage.--The first two blocks of information in Table 1 refer to "total file" and "50-State area only." The total file includes persons in covered employment in the 50-State area, abroad (including Puerto Rico and other outlying areas), and the military. The 50-State area excludes the military (which is treated in the file as a separate "State") and all persons working in covered employment outside territorial United States. Since the 50-State universe encompasses movement between States but excludes movement between the States and the military, we are able to isolate and focus on civilian interstate migration within the United States.

The third block of information contained in Table 1 refers to CPS-employed interstate migrants 18 to 64 years of age. The Social Security data include the entire working population, but for the purposes of this study, the File's population was considered to be synonymous with the age group 18 to 64. The ratios used to compare CPS employed interstate migrants with File migrants are therefore consistent as to area of movement and employment status, but not entirely consistent as to nature of the migration (change in place of employment versus place of residence) or age of migrants.

The migration base.--Of major substantive interest for this study is the observation that about 85 percent of persons in covered employment in a given year had also been working in the preceding year. These are the "matched cases" whose migration experience forms the core of this analysis. An 85 percent "match rate" is highly encouraging in itself; the absolute number of matched cases available from year to year produces a very substantial base for the computation of migration rates. 2/

Interstate migration and contiguity.--It is noteworthy that the Work History File generates a much larger number of interstate migrants than is found in the Current Population Surveys. (The overstatement is substantially greater when calendar-year data are used.) The excess is apparent in the Social Security data even though moves caused by interchange between the military and civilian employment have been excluded from the CWHS but not entirely from CPS. It is clear that the problem is caused primarily by changes in State of employment which are not accompanied by a change in residence. When the number of interstate movers from the two sources are compared separately for contiguous or noncontiguous States, the figures become more understandable.

If we examine moves between noncontiguous States, the number of migrants from the File does not differ greatly from the number obtained from the Current Population Survey, their ratios varying from 1.03 in 1962-63 to 1.25 in 1964-65 (first-quarter data). In the case of moves between contiguous States, the File generates from 35 percent to 55 percent more migrants than indicated by the Survey data. These data confirm what is evident, that many changes of employment between contiguous States do not involve any change in residence.

"Job migrants" versus resident migrants.--The problem of excess migrants being generated by the File is only one of several which arise when movements between contiguous States are examined. Indeed, excess migration would not be a matter for concern if the movement were proportionate to population and the balancing out resulted in a "true" net figure. This is not the case, since some States are in a more favorable position to gain "job migrants" as against resident migrants. Several examples were chosen to illustrate the directional bias involved. Migration stream data for 1955-60 (from 1960 Census data on residence migration) and for 1957-63 (from the Work History File) were assembled for a number of large metropolitan areas which form the nucleus of contiguous State movement. These include the New York metropolitan area (New York-New Jersey), Philadelphia (Pennsylvania-New Jersey), District of Columbia-Maryland-Virginia, and several others. The data are summarized below as follows:

Origin	Destination	CWHS (Social Security): All out-migrants from origin (1)	Out-migrants to indicated State (2)	Percent $(3)=\frac{(2)}{(1)} \times 100$
New York	New Jersey	647.3	111.8	17.3
New Jersey	New York	271.3	94.5	34.8
Pennsylvania	New Jersey	379.5	64.6	17.0
New Jersey	Pennsylvania	271.3	39.3	14.5
District of Col.	Maryland	93.3	28.4	30.4
Maryland	District of Col.	132.5	20.0	15.1
Rhode Island	Massachusetts	38.6	12.5	32.4
Massachusetts	Rhode Island	187.7	15.1	8.0
		Census Data: All out-migrants from origin (4)	**Out-migrants to indicated State (5)**	**Percent $(6)=\frac{(5)}{(4)} \times 100$**
New York	New Jersey	990.5	181.3	18.3
New Jersey	New York	388.5	74.4	19.1
Pennsylvania	New Jersey	678.6	115.0	17.0
New Jersey	Pennsylvania	388.5	57.6	14.8
District of Col.	Maryland	193.3	77.5	40.1
Maryland	District of Col.	259.7	18.3	7.0
Rhode Island	Massachusetts	86.1	17.8	20.7
Massachusetts	Rhode Island	339.6	16.5	4.9

It is clear that the overall impact of contiguous and noncontiguous State movements is substantial and must be considered in interpreting the migration data from the CWHS.

Regions, divisions, and States.--Because of differences in the geographic distribution of industries and occupations, the Work History Sample presents a biased view of migration by State of employment, which in turn compounds the difficulties of converting job migration into residence migration. States with larger proportions of their work force in covered employment have a disproportionate influence on migration derived from the File compared to States with smaller proportions.

Comparisons of the States' representation in the Work History File with the distribution of employed persons reported in the 1960 Census demonstrates the variation in worker coverage derived from the two sources. Table 2 presents ratios of one to the other. These ratios vary considerably from State to State, as expected. The lowest ratios are in the more rural Southern States and Plains States, notably North and South Dakota, Mississippi, Arkansas, and Iowa. In each of these States, the number of workers covered by Social Security provisions amounts to less than 70 percent of the working population counted by the Census. At the other end of the scale, there are industralized States like Delaware and New York whose ratios approach 100 percent.

As a result of these differences, the migration rates for States derived from the File are not of uniform validity as measures of residential migration of the total population. As the migration of covered workers represents only a portion of the migration of total employment, leaving a large uncovered category, the migration rates themselves may be biased simply because the opportunity of being reported as an out-migrant is higher in the high coverage States than in the low coverage States. Because of this differential exposure to risk, it appears more likely that the File will pick up in-migration to low coverage States from high coverage States, and less likely to reflect out-migration from low coverage States to high coverage States.

Evidence provided by the Current Population Surveys further complicates the picture by revealing sharp migration differences by occupation, even in occupations normally covered by Social Security provisions. The data in Table 1A, "Interstate Migration of the Employed Male Population 18-64 by Class of Worker and

Occupational Status: 1960-65" indicate that among wage and salary workers, farm workers had the highest interstate migration rate (5.5 percent), white collar workers the second (4.5 percent), and manual and service workers the lowest (2.8 percent). Thus, in States with the smallest proportions of workers in covered employment, there appears to be the strongest tendency toward interstate migration.

Interesting geographic differences by age and race are apparent from Table 3, in which workers in covered employment are compared with State populations 15-64 years of age in 1960. The two racial groups used in making this comparison are those available from the Work History File (which includes nonwhite races other than Negro) and Negro. Among the white population, Social Security workers comprised 58 percent of the national population 15-64 years of age in 1960. By State, the percentages were lowest in rural Southern and Plains States and highest in the industrialized Northeast, ranging from 43 percent in Kentucky to 69 percent in New York. Negro percentages, while about the same as the white nationally, are higher than white percentages in all States except Michigan and most of the Southern States. They also have a much wider spread than white percentages, extending from a low of 44 percent in Alabama to an astonishing high of 84 percent in Connecticut.

Net migration rates for regions and States.--Bearing in mind the limitations of the Social Security data, the net civilian migration rates derived from these data can now be compared with those taken from other sources. Tables 4, 5, and 6 present a variety of migration rates for regions and States which pertain to different periods. The rates were derived by dividing the net migrants cumulated for the 1960-65 period by the average annual matched workers in the File. Since only the persons we were able to match from year to year were exposed to the risk of migration, they were taken to constitute the appropriate population base for the computation of rates. Thus, persons in the File in only one of two successive years would not enter into the estimating equation at any time.

Table 4 shows closely comparable net migration rates by region. Civilian rates from the Work History File for 1960-65 are compared with rates for 1950-60, 1955-60, and 1960-65 taken from three different sources--the Censuses of 1950 and 1960, the Current Population Surveys, and independent estimates prepared at the Census Bureau. The Work History rates compare favorably with CPS rates, and less well with rates derived through independent estimates. They also suggest that 1960-65 regional net migration was very similar to that prevailing in the preceding 5-year period, 1955-60.

Differences for the South, however, are noteworthy. It has been speculated that the long-term net out-migration from the South may have halted in the 1960's, and that, in fact, there may have been a small net in-migration. The Work History Sample suggests, on the contrary, that net out-migration has continued at about the rate of the late 1950's. Indeed, looking at the data by race, it is not clear that a significant slowdown of net out-migration of Negroes from the South, which has been suggested, has been taking place at all. Independent estimates of net migration are not available by race, but both the Work History Sample and the CPS data suggest the continuation of net out-migration of Negroes from the South, roughly at the level of that of the late 1950's.

While the rate at which the white population is moving to the West appears to be confirmed by all three sources showing data by color, the Work History File shows a much heavier net in-migration of Negroes to the West in 1960-65 than does the CPS. Again, we must point to the nature of the coverage of the Work History File and note that the migration patterns of persons in covered employment may be atypical of the total population.

The net migration rates on a State-by-State basis (Table 5) provide some insight into the differences between the Work History File and the independent estimates for 1960-65. The differences are particularly acute in Florida and in the D.C.-Maryland-Virginia area. In both areas, peculiar local circumstances make utilization of an employment file highly questionable. Heavy net in-migration of the retired population to Florida would clearly not be reflected by the File. The D.C. Area presents a unique problem, for it is not a State, but the core of a large metropolitan area marked by heavy commuting from two States. In addition, its chief employer, the Federal Government, is hardly covered in the Work History File at all. Use of employment data for measuring residence changes is particularly inappropriate in this case.

The net migration rates obtained from the Work History File for the States more often than not agree with or are very close to the independent State estimates for the same periods. Differences in New York State may be attributed, in part, to the New York-New Jersey stream of migration and, in part, to the role played by net immigration from abroad. Although the effect of immigration from abroad is reflected in the Work History data as well, it is not possible to isolate this component for special study. As soon as an immigrant acquires a Social Security number and enters employment, he loses his original identity (as an immigrant) and is merged with all other workers in covered employment.

Table 6 shows net migration rates for States for white and Negro separately. These rates represent the only systematic measure of State Negro migration for the period since 1960. Unfortunately, the sampling errors on these rates appear to be quite large and their exact meaning has not yet been determined. However, the CWHS clearly indicates that many of the Southern States are still experiencing significant net out-migration of Negroes, many of them running well in excess of five percent.

Net migration rates, by sex, have also been computed. These appear in Appendix Table 1 where they are compared with 1955-60 residence migration rates obtained from the Census. Greater variation by sex is apparent in the Work History data than in the Census data. With only one exception, Hawaii, the Census migration rates by sex bear the same signs. Examples of opposing migration tendencies by sex are numerous in the Work History series; however, considerable differences in rates for males and females are also noticeable in some States, even when they are moving in the same direction.

Standard Metropolitan Statistical Areas

The manner in which the File is used for the derivation of migration patterns for SMSA's in the main resembles that already described for States. The size and selection of the sample, the use of first calendar quarters for measuring the migration period, the presumed age of the migrants (18-64), and the use of matched workers as the migration base are the same for SMSA's as for States. Analysis of Work History data for the SMSA's is likewise hampered by many of the same problems which affect the States. However, in the case of the SMSA's, there is the advantage that a change in Area of employment is more than likely to result in a change of Area of residence also.

The analysis for this paper was made on the basis of the 1968 definition of SMSA's. 1960 population figures were made to conform to the 1968 definition; new SMSA's added after April 1960 were included in the ranking and grouping which preceded the selection of SMSA's shown in Tables 7 and 8, and Appendix Table 2. All data drawn from the Work History File for 1960-65 similarly conform to the 1968 definition of Areas.

A given SMSA's migration is classified in two ways: as movement between the SMSA and other metropolitan areas, and as movement between the SMSA and nonmetropolitan areas. The military are shown as a separate component of nonmetropolitan area movement. By adding up the in- and out-migrants obtained for individual SMSA's, the metropolitan-nonmetropolitan exchange of workers in covered employment is closely approximated. Summary data rates computed for the country as a whole are shown below:

Total and Civilian Net Migration for SMSA's: 1960-65

(In thousands. 50-State area only.)

	Total 1/			Civilian 1/		
	All Classes	White 2/	Negro	All Classes	White 2/	Negro
Out-migrants from SMSA's						
Total	14,837.3	13,723.2	1,114.1	13,749.3	12,722.8	1,026.5
To other SMSA's ...	9,627.1	8,911.3	715.8	9,627.1	8,911.3	715.8
To non-SMSA's	5,210.2	4,811.9	398.3	4,122.2	3,811.5	310.7
In-migrants to SMSA's						
Total	15,503.1	14,314.3	1,188.8	14,110.9	13,011.7	1,099.2
From other SMSA's..	9,627.1	8,911.3	715.8	9,627.1	8,911.3	715.8
From non-SMSA's....	5,876.0	5,403.0	473.0	4,483.8	4,100.4	383.4
Net migrants to SMSA's from non-SMSA's	+665.8	+591.1	+74.7	+361.6	+288.9	+72.7
Percent net migrants of base population						
SMSA's	+2.0	+1.9	+2.4	+1.1 3/	+0.9 3/	+2.3 3/
Non-SMSA's	-3.9	-3.9	-4.3	-2.1	-1.9	-4.2

1/ "Total" includes moves to and from military; "civilian" excludes military moves.

2/ Includes nonwhite population other than Negro.

3/ Base includes military.

Unlike the situation noted for States, the military component of SMSA migration was identified, but was not deleted from the total matched work force. Pure civilian migration rates could thus not be computed for SMSA's in the above text table. For the individual SMSA's shown in Table 8, the net migration rates include movement to and from the military and can be compared directly with the 1960-65 rates from the Bureau's independent estimates which also include the military.

In compiling SMSA data for this paper, we restricted our universe to the 55 metropolitan areas whose populations in 1960 included 50,000 or more Negroes. These SMSA's were subdivided into three groups ranked by the number of Negroes in 1960. Table 7 shows for these 55 Areas ratios of workers from the Social Security File to 1960 Census population 15-64 years of age by race. Surprisingly, the percent workers of population varies widely among the Areas for both races: for whites, from a low of 30 to a high of 69; for nonwhites from a low of 32 to a high of 64. One would expect somewhat less variation in coverage among metropolitan population than for States.

The question of how well statistics from the Work History Sample reflect the race and age composition of SMSA's may be answered, in part, by the comparison shown in Table 2 of the Appendix. In this table, we see the percent Negro of the 1960 Census population 15-64 years of age alongside the percent Negro of the total Work History File in 1960. The two columns are remarkably similar for most SMSA's. All except four SMSA's are within five percentage points' difference of one another. 44 are within three percentage points of one another.

Net migration rates for SMSA's.--Table 8 contrasts net migration rates produced by the Work History File with rates taken from the Bureau's independent estimates. The latter refer to residential migration and to all ages only. They are not available by race. Overall by size class, the net migration rates derived from these two sources compare very favorably. Considered individually, in 43 of the 55 SMSA's shown, there is agreement on the basic question of whether there was a net gain or a net loss of population through migration between 1960 and 1965. (Of the 12 which do not agree, 10 are in the South, and 3 of these are Florida resort or retirement centers.) In more than half of the SMSA's there is substantial correspondence (i.e., less than three percentage points difference) between the rates obtained from the independent estimates and rates yielded by the Social Security data. There is no discernible geographic or size pattern which would account for the fact that estimates for some SMSA's compare more favorably than for others or which explains the several very large differences.

Sampling errors

A thorough investigation of the effect of sampling error on the migration rates shown in the various tables is still in process.3/ A preliminary review suggests that the File will not provide very meaningful rates (i.e., not statistically different from zero) on an annual basis. Computed over a longer period of time, however, such as the five-year period used here, the sampling errors become manageable in a large number of instances. Rates for regions shown in Table 4 have already been tested and found significant (2-sigma level) in most cases. Rates for the large States and those generating large numbers of migrants (e.g., Negroes of Southern States) are also likely to be significant. This is probably true for the large SMSA's also. Rates for small States and SMSA's may fail the test of significance. (See Appendix Tables 3 and 4.)

Conclusion

The Continuous Work History Sample can be manipulated to provide postcensal estimates of net migration (resident) at a level of demographic and geographic detail which has not hitherto been available. There are, however, serious limitations to the use and interpretation of CWHS-derived migration rates, particularly when considered in conjunction with their use in a regular program of postcensal population estimates such as that carried on by the Bureau of the Census. What we see as the favorable aspects of the File for measurement of resident migration are:

1. The broad extent of Social Security coverage, amounting to 90 percent of total civilian employment.

2. The high percentage of matched cases from year to year, i.e., of workers whose migration experience we are able to follow, accounting for 85 percent of the total Social Security File in the 1960-65 period.

3. Consistent annual data on a first-quarter basis from 1963 on, which allows migration to be measured for fairly precise time intervals.

4. The race detail available from the File which compares well with Census population data for both States and SMSA's (see Table 3 and Appendix Table 2).

5. The fair degree of consistency in net migration rates computed from different data sources for the 1955-60 and 1960-65 periods (see tables 4 and 6).

On the negative side, there are still several problems to consider:

1. We must point out again that in spite of the apparent consistency between the CWHS net migration rates and those from other sources noted above, we have not established the extent to which migration rates based on employment changes reflect true resident migration. Here, we need to wait for the 1965-70 resident migration data which will be forthcoming from the 1970 Census to provide a firmer basis for analysis than is now available.

2. The size of the sample, which probably precludes deriving rates for many States and SMSA's in which we are interested, even cumulated over a number of years; and

3. The timeliness of the data. The timing of the CWHS has to be substantially improved if migration rates derived from the File are to be useful in any <u>regular</u> current program of population estimation. The lag is now close to three years, for the last year for which data are available is 1966.

Footnotes

1/ <u>Current Population Reports</u>, Series P-20, annual report on mobility, of which No. 171, "Mobility of the Population of the United States, March 1966 to March 1967," is the most recent issue.

2/ Although the number of covered workers not matched in the File is not considered in this study, it should be of interest to those concerned with gross changes in the labor force. Looking ahead from year to year, unmatched cases in the 50-State area represent mainly persons who leave the labor force (or the 50-State area), shift to "noncovered" employment, enter the military, or die. Looking backward, new entrants (including persons returning from military duty) probably make up the bulk of the group.

3/ Investigation of the effect of sampling error on the CWHS-derived migration rates shown in this paper was completed after the main text had been written. The results of this investigation are shown in Appendix Tables 3 and 4 which contain standard errors (1 Sigma) of net migration rates by color for States and SMSA's. The results generally support conclusions drawn from a preliminary review of the material.

Table 1.--SUMMARY DATA FROM CONTINUOUS WORK HISTORY SAMPLE OF SOCIAL SECURITY AND COMPARISON OF INTERSTATE MIGRANTS FROM CWHS AND CPS: 1961 TO 1965

(Numbers in thousands)

	Calendar Year 1/				First Quarter of Year 2/		
	1961	1962	1963	1964	1963	1964	1965
Total file 3/							
Working in year indicated	66,434.7	68,058.4	69,466.7	71,474.0	56,591.0	57,854.9	60,078.2
Working in preceding year	59,594.7	60,515.4	61,884.5	63,415.9	48,610.9	49,684.1	50,677.5
Same State	55,148.3	56,082.4	57,286.5	58,675.9	45,346.3	46,402.3	47,205.0
Different State - number	4,446.4	4,433.0	4,598.0	4,740.0	3,264.6	3,281.8	3,472.5
- percent	7.5	7.3	7.4	7.5	6.7	6.6	6.9
50-State area only 4/							
Working in year indicated	62,969.0	64,358.5	65,676.6	67,681.7	53,351.1	54,462.8	56,431.2
Working in preceding year	55,922.7	56,643.0	57,825.0	59,272.1	45,233.9	46,259.8	48,154.3
Same State	52,498.6	53,240.7	54,319.2	55,712.6	42,871.2	43,876.2	45,650.8
Different State - number	3,424.1	3,402.3	3,505.8	3,559.5	2,362.7	2,383.6	2,503.5
- percent	6.1	6.0	6.1	6.0	5.2	5.2	5.2
Contiguous State	1,439.2	1,450.7	1,463.5	1,492.9	994.2	1,012.4	1,080.3
Noncontiguous State	1,984.9	1,951.6	2,042.3	2,066.6	1,368.5	1,371.2	1,423.2
Interstate migrants from CPS employed persons only 5/							
Total 18-64 - number	1,778.0	1,702.0	2,052.0	1,894.0	2,052.0	1,894.0	1,839.0
- percent	3.0	2.8	3.3	3.0	3.3	3.0	2.8
Contiguous State	NA	616.0	725.0	677.0	725.0	677.0	697.0
Noncontiguous State	NA	1,086.0	1,327.0	1,217.0	1,327.0	1,217.0	1,142.0
Ratio: CWHS interstate migrants to CPS							
Total	1.93	2.00	1.71	1.88	1.15	1.26	1.36
Contiguous State	NA	2.36	2.02	2.21	1.37	1.50	1.55
Noncontiguous State	NA	1.80	1.54	1.70	1.03	1.13	1.25

1/ Includes persons who worked at any time during year.

2/ Includes only those working in first calendar quarter.

3/ Includes persons working in the 50 States, U.S. territories and possessions, on ships at sea, and military personnel.

4/ Excludes military personnel and all persons working outside the 50 States.

5/ Employed at time of Current Population Survey. Excludes persons in Armed Forces.

NA - Not available.

Table 1A.--INTERSTATE MIGRATION OF THE EMPLOYED MALE POPULATION 18-64 BY CLASS OF WORKER AND OCCUPATIONAL STATUS: CURRENT POPULATION SURVEY, 1960-65 AVERAGE

	Interstate Migration Rate (percent)
Total male civilian population 18-64	3.5
Employed	3.1
Wage and salary 1/	3.5
White collar workers	4.5
Manual and service workers	2.8
Farm workers	5.5
Self-employed	1.2

1/ Includes government workers.

Table 2.--COMPARISON OF CWHS FILE AND CENSUS: 1960

(Numbers in thousands)

Region, Division, and State	Workers		
	CWHS 1960 1/	Census 1960 2/	CWHS as a percent of Census
United States, Total	62,715.8	75,634.7	82.9
Regions			
Northeastern States	17,969.3	19,662.7	91.4
North Central States	18,043.2	22,055.7	81.8
The South	17,018.0	21,934.8	77.6
The West	9,685.3	11,981.5	80.8
Northeast			
New England	4,072.7	4,706.0	86.5
Middle Atlantic	13,896.6	14,956.6	92.9
North Central			
East North Central	13,095.1	15,452.3	84.7
West North Central	4,948.1	6,603.4	74.9
South			
South Atlantic	8,547.4	10,572.4	80.8
East South Central	3,354.1	4,655.1	72.1
West South Central	5,116.5	6,707.3	76.3
West			
Mountain	2,225.5	2,802.0	79.4
Pacific	7,459.8	9,179.5	81.3
New England			
Maine	332.0	417.9	79.4
New Hampshire	244.7	279.3	87.6
Vermont	125.7	172.3	73.0
Massachusetts	2,015.6	2,314.9	87.1
Rhode Island	321.0	369.3	86.9
Connecticut	1,033.7	1,152.3	89.7
Middle Atlantic			
New York	7,318.0	7,539.2	97.1
New Jersey	2,353.4	2,659.8	88.5
Pennsylvania	4,225.2	4,757.6	88.8
East North Central			
Ohio	3,385.8	4,057.1	83.5
Indiana	1,671.3	1,989.9	84.0
Illinois	3,861.6	4,472.8	86.3
Michigan	2,807.4	3,207.7	87.5
Wisconsin	1,369.0	1,724.8	79.4
West North Central			
Minnesota	1,077.7	1,477.8	72.9
Iowa	836.8	1,204.1	69.5
Missouri	1,568.2	1,832.0	85.6
North Dakota	157.4	266.7	59.0
South Dakota	181.1	285.4	63.5
Nebraska	463.6	613.4	75.6
Kansas	663.3	924.0	71.8

Table 2 continued

Region, Division, and State	Workers		
	CWHS 1960 1/	Census 1960 2/	CWHS as a percent of Census
South Atlantic			
Delaware	185.2	187.8	98.6
Maryland	1,049.9	1,296.5	81.0
District of Columbia	398.5	379.8	104.9
Virginia	1,211.9	1,561.6	77.6
West Virginia	530.4	642.2	82.6
North Carolina	1,517.8	1,936.9	78.4
South Carolina	705.4	941.4	74.9
Georgia	1,305.3	1,609.4	81.1
Florida	1,643.0	2,016.8	81.5
East South Central			
Kentucky	775.3	1,100.2	70.5
Tennessee	1,124.9	1,443.4	77.9
Alabama	914.2	1,255.0	72.8
Mississippi	539.7	856.5	63.0
West South Central			
Arkansas	460.5	716.8	64.2
Louisiana	880.0	1,185.5	74.2
Oklahoma	718.0	932.2	77.0
Texas	3,058.0	3,872.8	79.0
Mountain			
Montana	219.4	285.9	76.7
Idaho	210.1	289.0	72.7
Wyoming	113.4	144.8	78.3
Colorado	563.6	741.3	76.0
New Mexico	261.7	338.9	77.2
Arizona	430.0	507.8	84.7
Utah	309.5	362.1	85.5
Nevada	117.8	132.3	89.0
Pacific			
Washington	972.8	1,228.7	79.2
Oregon	630.2	804.2	78.4
California	5,589.7	6,827.2	81.9
Alaska	60.0	79.2	75.8
Hawaii	207.1	240.2	86.2

1/ Excluding Armed Forces.

2/ Civilian population 14 years of age and over who worked at all in 1959.

Table 3.--CWHS FILE WORKERS COMPARED WITH 1960 CENSUS POPULATION 15-64 YEARS OF AGE, BY RACE

(Numbers in thousands)

Region, Division, and State	White 1/		Negro	
	CWHS 1960 2/	Percent CWHS of Population 15-64 Years	CWHS 1960 2/	Percent CWHS of Population 15-64 Years
United States, Total...	56,466.0	58.4	6,249.8	59.0
Regions				
Northeastern States	16,549.8	64.5	1,419.5	75.9
North Central States	16,792.1	59.0	1,251.1	63.0
The South	13,821.0	52.3	3,197.0	52.5
The West	9,303.1	57.5	382.2	59.3
Northeast				
New England	3,960.1	63.9	112.6	78.7
Middle Atlantic	12,589.7	64.7	1,306.9	75.7
North Central				
East North Central	12,054.3	60.7	1,040.8	62.2
West North Central	4,737.8	55.0	210.3	67.2
South				
South Atlantic	6,746.8	54.7	1,800.6	56.0
East South Central	2,721.7	48.4	632.4	45.1
West South Central	4,352.5	51.3	764.0	51.8
West				
Mountain	2,179.2	56.0	46.3	65.6
Pacific	7,123.9	58.0	335.9	58.5
New England				
Maine	330.9	58.9	(Z)	(B)
New Hampshire	243.9	68.2	(Z)	(B)
Vermont	125.2	55.8	(Z)	(B)
Massachusetts	1,964.9	64.6	50.7	77.2
Rhode Island	314.6	60.7	(Z)	(B)
Connecticut	980.6	65.9	53.1	83.9
Middle Atlantic				
New York	6,582.7	68.6	735.3	81.4
New Jersey	2,104.2	60.8	249.2	80.3
Pennsylvania	3,902.8	60.9	322.4	63.0
East North Central				
Ohio	3,111.4	58.8	274.4	59.9
Indiana	1,570.3	60.5	101.0	66.3
Illinois	3,465.7	62.8	395.9	65.5
Michigan	2,570.0	61.4	237.4	56.7
Wisconsin	1,336.9	59.4	(Z)	(B)
West North Central				
Minnesota	1,065.1	55.0	(Z)	(B)
Iowa	825.9	52.7	(Z)	(B)
Missouri	1,425.4	60.8	142.8	65.4
North Dakota	156.7	43.8	(Z)	(B)
South Dakota	180.7	47.3	(Z)	(B)
Nebraska	450.4	56.5	(Z)	(B)
Kansas	633.6	51.8	29.7	59.0

Table 3 continued

Region, Division, and State	White 1/		Negro	
	CWHS 1960 2/	Percent CWHS of Population 15-64 Years	CWHS 1960 2/	Percent CWHS of Population 15-64 Years
South Atlantic				
Delaware	157.1	67.4	28.1	80.3
Maryland	822.4	51.8	227.5	76.0
District of Columbia	287.1	115.2	111.4	44.0
Virginia	944.8	48.5	267.1	58.2
West Virginia	504.7	48.2	(Z)	(B)
North Carolina	1,204.5	56.4	313.3	52.3
South Carolina	513.1	53.3	192.3	45.4
Georgia	989.9	57.2	315.4	52.5
Florida	1,323.2	54.2	319.8	64.2
East South Central				
Kentucky	703.2	42.6	72.1	59.5
Tennessee	955.8	52.6	169.1	52.8
Alabama	692.0	49.8	222.2	43.5
Mississippi	370.7	48.8	169.0	37.6
West South Central				
Arkansas	374.5	45.0	86.0	44.6
Louisiana	630.4	47.4	249.6	46.2
Oklahoma	674.5	51.7	43.5	53.7
Texas	2,673.1	53.3	384.9	58.2
Mountain				
Montana	219.0	57.2	(Z)	(B)
Idaho	209.4	55.4	(Z)	(B)
Wyoming	112.8	58.7	(Z)	(B)
Colorado	547.6	54.5	(Z)	(B)
New Mexico	256.0	48.2	(Z)	(B)
Arizona	415.3	56.3	(Z)	(B)
Utah	306.7	61.8	(Z)	(B)
Nevada	112.4	65.4	(Z)	(B)
Pacific				
Washington	957.5	57.8	(Z)	(B)
Oregon	623.0	60.3	(Z)	(B)
California	5,277.7	58.2	312.0	59.2
Alaska	59.2	43.7	(Z)	(B)
Hawaii	206.5	53.9	(Z)	(B)

1/ Includes nonwhite races other than Negro.

2/ Excludes Armed Forces.

(Z) - Having less than 50,000 Negro population 15-64 years of age in 1960.

(B) - Base less than 50,000.

Table 4.--NET MIGRATION RATES FOR SELECTED PERIODS, BY REGION AND COLOR

Region and Color	1950-60 1/ Censuses of 1950 and 1960	1955-60 2/ Census of 1960	1960-65 3/ CWHS Data	1960-65 4/ Current Population Surveys	1960-65 5/ Independent estimate by Census Bureau
U. S., All Classes	+1.8	-	-	-	+1.1
Northeastern States	+0.9	-1.7	-0.4	-0.8	+1.2
North Central States	-0.3	-1.9	-1.6	-2.2	-1.8
The South	-3.0	-0.1	-0.7	-1.1	+1.3
The West	+19.1	+6.2	+4.8	+6.9	+5.6
U. S., White	+2.0	-	-	-	
Northeastern States	-0.6	-2.1	-0.9	-1.5	
North Central States	-1.6	-2.2	-1.8	-2.6	
The South	+0.1	+0.7	+0.2	-0.3	
The West	+18.7	+6.2	+4.5	+7.1	
U. S., Nonwhite	-0.2	-	-	-	
Northeastern States	+26.0	+3.8	+6.1	+6.4	
North Central States	+23.8	+2.3	+1.4	+1.9	
The South	-14.1	-3.4	-4.9	-4.2	
The West	+23.6	+6.5	+11.8	+5.6	

1/ All ages. Includes military and immigrants from abroad.

2/ Population 15-64 years of age in 1960. Includes military.

3/ 18-64 years of age. Includes an unknown number of immigrants from abroad. Excludes military. "Nonwhite" here refers to Negro only.

4/ Population 1 year of age and over. Excludes Armed Forces in barracks.

5/ All ages. Includes immigrants from abroad. Excludes military.

Table 5.--NET MIGRATION RATES FOR STATES AND REGIONS FROM CENSUS, CWHS, AND INDEPENDENT ESTIMATE: 1955-60 AND 1960-65

Region, Division, and State	1955-60 1/ Census	1960-65 2/ CWHS	1960-65 3/ Independent Estimate
United States, Total	-	-	+1.1
Regions			
Northeastern States	-1.7	-0.4	+1.2
North Central States	-1.9	-1.6	-1.8
The South	-0.1	-0.7	+1.3
The West	+6.2	+4.8	+5.6
Northeast			
New England	-0.7	-0.2	+0.2
Middle Atlantic	-1.9	-0.4	+1.5
North Central			
East North Central	-1.3	-1.3	-1.2
West North Central	-3.2	-2.3	-3.3
South			
South Atlantic	+2.3	-0.6	+2.7
East South Central	-3.4	-1.6	-1.0
West South Central	-1.3	-0.2	+0.7
West			
Mountain	+4.4	+1.5	+2.3
Pacific	+6.8	+5.7	+6.7
New England			
Maine	-3.4	-5.4	-4.7
New Hampshire	+1.4	-0.7	+4.8
Vermont	-2.8	-0.7	-2.8
Massachusetts	-1.4	-0.1	-1.5
Rhode Island	-1.6	-1.2	-1.4
Connecticut	+1.7	+1.7	+5.3
Middle Atlantic			
New York	-2.6	-1.9	+2.5
New Jersey	+2.2	+7.9	+5.9
Pennsylvania	-3.2	-2.5	-2.4
East North Central			
Ohio	-0.6	-2.5	-0.8
Indiana	-1.2	(Z)	-1.9
Illinois	-1.2	-0.4	-0.9
Michigan	-2.6	-1.3	-1.2
Wisconsin	-1.2	-2.4	-2.2
West North Central			
Minnesota	-1.1	-0.7	-3.3
Iowa	-4.5	-6.0	-5.8
Missouri	-1.8	-2.3	-1.1
North Dakota	-7.2	-7.2	-6.7
South Dakota	-7.0	-6.2	-6.9
Nebraska	-5.1	-3.8	-3.5
Kansas	-4.1	+2.8	-2.5

Table 5 continued

Region, Division, and State	1955-60 1/ Census	1960-65 2/ CWHS	1960-65 3/ Independent Estimate
South Atlantic			
Delaware	+4.7	+9.2	+4.8
Maryland	+3.6	+7.2	+5.5
District of Columbia	-10.1	-19.6	-2.8
Virginia	+1.8	+2.4	+3.1
West Virginia	-9.8	-5.6	-7.9
North Carolina	-2.1	-2.3	-0.1
South Carolina	-2.2	+0.3	-1.8
Georgia	-1.2	-3.3	+2.1
Florida	+17.2	+0.4	+9.9
East South Central			
Kentucky	-4.7	-0.7	-2.3
Tennessee	-2.9	-0.4	+0.7
Alabama	-2.3	-2.7	-0.9
Mississippi	-4.4	-3.4	-2.1
West South Central			
Arkansas	-5.4	+1.0	+1.4
Louisiana	-0.4	+1.8	-0.6
Oklahoma	-4.2	-1.6	-0.6
Texas	-0.2	-0.7	+1.2
Mountain			
Montana	-3.1	-5.5	-4.0
Idaho	-2.4	-7.1	-3.9
Wyoming	-2.1	-8.4	-8.6
Colorado	+4.1	+3.5	+2.9
New Mexico	+5.7	-10.5	-5.7
Arizona	+14.0	+8.0	+9.3
Utah	+0.9	-0.2	-0.3
Nevada	+9.2	+28.7	+32.9
Pacific			
Washington	+1.3	+1.2	-1.8
Oregon	-0.7	+0.4	+3.8
California	+8.7	+7.2	+8.7
Alaska	+8.3	+10.1	+3.7
Hawaii	+2.7	-0.4	+1.0

1/ Net migrants as a percent of census population 15-64 years of age. Includes military.

2/ Net employed civilian migrants as a percent of average base population in Work History File. 1960-65 includes unknown number of immigrants from abroad.

3/ Net civilian migrants as percent of mid-period population. Includes immigrants from abroad.

(Z) - Less than 0.05.

Table 6.--NET MIGRATION RATES BY COLOR FROM CENSUS AND CWHS DATA: SELECTED PERIODS, 1950-1965

Region, Division, and State	White			Nonwhite		
	1950-60 1/ Censuses	1955-60 2/ Census	1960-65 3/ CWHS	1950-60 1/ Censuses	1955-60 2/ Census	1960-65 3/ CWHS
United States, Total	+2.0	-	-	-0.2	-	-
Regions						
Northeastern States....	-0.6	-2.1	-0.9	+26.0	+3.8	+6.1
North Central States...	-1.6	-2.2	-1.8	+23.8	+2.3	+1.4
The South	+0.1	+0.7	+0.2	-14.1	-3.4	-4.9
The West	+18.7	+6.2	+4.5	+23.6	+6.5	+11.8
Northeast						
New England	-0.5	-0.9	-0.5	+45.6	+8.8	+12.6
Middle Atlantic	-0.6	-2.5	-1.0	+24.5	+3.4	+5.6
North Central						
East North Central	+0.6	-1.7	-1.5	+28.1	+2.5	+1.9
West North Central	-6.3	-3.4	-2.4	+7.6	+1.3	-0.7
South						
South Atlantic	+7.4	+3.5	+0.4	-10.5	-2.4	-4.5
East South Central	-9.6	-2.8	-0.2	-22.9	-6.2	-8.2
West South Central	-2.4	-1.0	+0.3	-11.8	-2.9	-3.3
West						
Mountain	+11.3	+4.5	+1.5	+3.6	+3.1	+2.7
Pacific	+21.3	+6.7	+5.4	+27.4	+7.1	+12.9
New England						
Maine	-7.5	-3.5	-5.4	(B)	(B)	(B)
New Hampshire	+2.1	+1.2	-0.4	(B)	(B)	(B)
Vermont	-10.1	-2.8	-0.9	(B)	(B)	(B)
Massachusetts	-2.6	-1.6	-0.5	+32.1	+6.7	+12.1
Rhode Island	-3.6	-1.8	-1.5	(B)	(B)	(B)
Connecticut	+10.0	+1.4	+1.0	+71.1	+10.3	+14.1
Middle Atlantic						
New York	-0.5	-3.3	-2.5	+29.5	+3.9	+4.3
New Jersey	+10.3	+1.8	+7.0	+34.6	+6.5	+15.9
Pennsylvania	-5.6	-3.6	-2.7	+12.0	+0.7	+0.6
East North Central						
Ohio	+3.7	-0.9	-2.7	+25.6	+3.1	+0.3
Indiana	+0.5	-1.4	-0.1	+25.4	+2.5	+1.3
Illinois	-0.8	-1.7	-0.9	+28.3	+3.2	+4.3
Michigan	+0.5	-2.9	-1.4	+27.9	(Z)	+0.1
Wisconsin	-2.4	-1.4	-2.5	(B)	(B)	(B)
West North Central						
Minnesota	-3.4	-1.2	-0.9	(B)	(B)	(B)
Iowa	-9.1	-4.6	-6.1	(B)	(B)	(B)
Missouri	-4.3	-2.1	-2.0	+9.3	+1.0	-4.8
North Dakota	-16.9	-7.3	-6.6	(B)	(B)	(B)
South Dakota	-14.3	-7.3	-6.4	(B)	(B)	(B)
Nebraska	-9.3	-5.4	-3.7	(B)	(B)	(B)
Kansas	-2.7	-4.3	+2.1	+6.5	+0.2	+16.7

Table 6 continued

Region, Division, and State	White 1950-60 1/ Censuses	White 1955-60 2/ Census	White 1960-65 3/ CWHS	Nonwhite 1950-60 1/ Censuses	Nonwhite 1955-60 2/ Census	Nonwhite 1960-65 3/ CWHS
South Atlantic						
Delaware	+21.0	+4.9	+10.7	+14.6	+3.5	-0.5
Maryland	+14.5	+3.9	+6.6	+9.3	+2.2	+9.8
District of Columbia...	-41.1	-25.9	-19.0	+19.2	+4.6	-21.3
Virginia	+3.3	+2.9	+3.2	-9.5	-2.9	-0.3
West Virginia	-21.5	-9.7	-5.5	(B)	(B)	(B)
North Carolina	-4.0	-1.1	-0.6	-19.2	-5.6	-9.8
South Carolina	-0.3	+0.7	+2.2	-26.5	-8.6	-5.6
Georgia	-0.4	+0.1	-1.9	-19.2	-5.0	-8.0
Florida	+70.0	+19.8	+1.2	+16.6	+4.7	-3.3
East South Central						
Kentucky	-13.7	-4.9	-0.7	-7.6	-3.3	-1.1
Tennessee	-7.8	-2.8	+0.2	-10.7	-3.0	-3.7
Alabama	-6.9	-0.9	+0.1	-22.8	-6.0	-12.4
Mississippi	-9.3	-1.4	-0.7	-32.7	-9.4	-10.5
West South Central						
Arkansas	-19.1	-4.5	+2.9	-35.0	-9.3	-8.3
Louisiana	+2.4	+0.7	+4.7	-10.4	-3.0	-6.5
Oklahoma	-9.5	-4.2	-1.5	-13.0	-3.9	-2.9
Texas	+2.1	-0.1	-0.7	-2.7	-0.8	-0.3
Mountain						
Montana	-4.0	-3.2	-5.6	(B)	(B)	(B)
Idaho	-7.0	-2.4	-7.0	(B)	(B)	(B)
Wyoming	-6.5	-2.0	-8.0	(B)	(B)	(B)
Colorado	+11.5	+4.0	+3.6	(B)	(B)	(B)
New Mexico	+8.5	+6.0	-10.8	(B)	(B)	(B)
Arizona	+51.9	+15.3	+8.2	(B)	(B)	(B)
Utah	+1.4	+0.9	(Z)	(B)	(B)	(B)
Nevada	+53.2	+9.0	+28.8	(B)	(B)	(B)
Pacific						
Washington	+3.0	+1.2	+1.2	(B)	(B)	(B)
Oregon	+0.7	-0.7	+0.1	(B)	(B)	(B)
California	+28.2	+8.5	+6.8	+52.7	+11.0	+13.1
Alaska	+45.5	+9.1	+9.7	(B)	(B)	(B)
Hawaii	+48.0	+15.2	-0.2	(B)	(B)	(B)

1/ All ages. Includes military and immigrants from abroad.

2/ Net migrants as a percent of census population 15-64 years of age. Includes military.

3/ Net employed civilian migrants as a percent of average base population in the Work History File, 1960-65. Includes an unknown number of immigrants from abroad. "Nonwhite" here refers to Negro only.

(Z) Less than 0.05.

(B) Having less than 50,000 nonwhite or Negro population 15-64 years of age in 1960.

Table 7.--RATIO OF WORKERS FROM CWHS FILE TO CENSUS POPULATION 15-64 YEARS OF AGE, BY RACE, FOR STANDARD METROPOLITAN STATISTICAL AREAS WITH 50,000 OR MORE NEGRO POPULATION IN 1960

Standard Metropolitan Statistical Area 1/	Ratio of Workers From CWHS, 1960 to Census Population 15-64 Years of Age	
	White 2/	Negro
250,000+ Negro population, 1960		
New York, N. Y.	60.3	60.5
Los Angeles-Long Beach, Calif.	54.7	48.3
Chicago, Ill.	58.9	50.4
Philadelphia, Pa.-N. J.	52.5	47.3
Detroit, Mich.	56.5	42.9
St. Louis, Mo.-Ill.	55.5	45.9
Washington, D. C.-Md.-Va.	40.4	43.5
Cleveland, Ohio	57.2	50.6
Baltimore, Md.	52.2	50.3
Houston, Texas	50.7	44.6
New Orleans, La.	47.9	43.5
Memphis, Tenn.-Ark.	49.7	37.1
100,000-249,000 Negro population, 1960		
San Francisco-Oakland, Calif.	49.0	33.8
Pittsburgh, Pa.	51.1	41.4
Newark, N. J.	58.1	47.9
Cincinnati, Ohio-Ky.-Ind.	52.5	42.1
Dallas, Texas	60.5	52.5
Kansas City, Mo.-Kans.	56.0	49.0
Atlanta, Ga.	65.5	52.8
Indianapolis, Ind.	60.9	48.1
Miami, Fla.	54.8	64.2
Birmingham, Ala.	52.1	41.6
Norfolk-Portsmouth, Va.	29.7	42.5
Greensboro-Winston-Salem-High Point, N.C.	66.5	56.3
Jacksonville, Fla.	42.6	44.5
Richmond, Va.	65.8	60.4
Mobile, Ala.	38.2	32.4
50,000-99,000 Negro population, 1960		
Boston, Mass.	66.5	60.5
Buffalo, N. Y.	54.8	51.8
Milwaukee, Wisc.	57.9	59.2
Tampa-St. Petersburg, Fla.	44.0	44.5
Columbus, Ohio	50.0	38.2
Dayton, Ohio	50.8	36.5
Louisville, Ky.-Ind.	52.9	49.8
Gary-Hammond-East Chicago, Ind.	58.3	48.4
Fort Worth, Texas	47.5	42.9
Nashville, Tenn.	55.6	38.0
Fort Lauderdale-Hollywood, Fla.	39.7	48.0
Orlando, Fla.	44.9	38.5
Charlotte, N. C.	69.0	62.1
Beaumont-Port Arthur-Orange, Texas	47.5	43.5
Chattanooga, Tenn.-Ga.	61.2	48.9
Shreveport, La.	38.3	35.4
Little Rock-North Little Rock, Ark.	50.7	47.9
Columbia, S. C.	41.9	43.0
Charleston, S. C.	34.5	37.7
Baton Rouge, La.	44.2	35.9
West Palm Beach, Fla.	47.1	36.9
Newport News-Hampton, Va.	34.2	37.9
Jackson, Miss.	61.0	37.6
Columbus, Ga.-Ala.	36.5	37.0
Augusta, Ga.-S.C.	41.6	36.9
Montgomery, Ala.	55.1	39.3
Savannah, Ga.	50.2	48.0
Macon, Ga.	44.5	43.6

1/ As defined in 1968. 2/ Includes nonwhite races other than Negro.

Table 8.--NET MIGRATION RATES BY RACE FOR SMSA'S WITH 50,000 OR MORE NEGRO POPULATION IN 1960: SOCIAL SECURITY DATA AND INDEPENDENT ESTIMATES FOR 1960-65

Standard Metropolitan Statistical Area Rank and Size Group	Social Security Data 1/			Independent Estimate 2/
	All Classes	White 3/	Negro	All Classes
250,000+ Negro population, 1960				
New York, N. Y.	+0.3	-0.5	+6.5	+1.1
Los Angeles-Long Beach, Calif.	+10.3	+9.7	+18.6	+4.7
Chicago, Ill.	+0.2	-0.2	+3.5	+0.4
Philadelphia, Pa.-N. J.	-2.3	-2.6	-0.3	+1.3
Detroit, Mich.	-0.2	-0.7	+3.5	-1.4
St. Louis, Mo.-Ill.	+3.0	+3.2	+2.0	-0.1
Washington, D. C.-Md.-Va.	+10.2	+10.5	+9.3	+10.1
Cleveland, Ohio	-0.5	-0.6	-0.2	-1.5
Baltimore, Md.	-0.3	-0.4	-0.1	+0.2
Houston, Texas	+6.6	+7.5	+2.2	+8.9
New Orleans, La.	+10.8	+13.8	+2.6	+4.6
Memphis, Tenn.-Ark.	+1.7	+3.2	-2.3	+0.7
100,000-249,000 Negro population, 1960				
San Francisco-Oakland, Calif.	+6.8	+6.5	+12.2	+3.7
Pittsburgh, Pa.	-0.7	-0.4	-6.5	-6.0
Newark, N. J.	+3.4	+2.6	+10.5	+3.8
Cincinnati, Ohio-Ky.-Ind.	-3.8	-4.5	+4.0	-1.1
Dallas, Texas	+6.5	+6.6	+6.0	+8.8
Kansas City, Mo.-Kans.	+4.3	+4.5	+2.3	+0.8
Atlanta, Ga.	-0.4	+0.4	-4.6	+9.5
Indianapolis, Ind.	-0.2	-0.1	-2.0	-0.8
Miami, Fla.	-7.8	-7.7	-8.1	+7.6
Birmingham, Ala.	-7.3	-6.4	-10.6	-4.7
Norfolk-Portsmouth, Va.	-0.5	+4.0	-11.1	-0.3
Greensboro-Winston-Salem-High Point, N.C.	+3.1	+4.2	-2.7	+0.4
Jacksonville, Fla.	+4.1	+5.6	-1.9	-0.2
Richmond, Va.	+3.8	+6.2	-4.4	+4.3
Mobile, Ala.	-1.9	-0.2	-7.4	-2.1
50,000-99,000 Negro population, 1960				
Boston, Mass.	(Z)	-0.4	+14.0	-2.7
Buffalo, N. Y.	-2.2	-2.7	+5.6	-4.9
Milwaukee, Wisc.	-1.5	-1.9	+6.7	-4.2
Tampa-St. Petersburg, Fla.	-2.1	-1.7	-4.7	+9.4
Columbus, Ohio	+1.8	+2.0	-0.7	+3.4
Dayton, Ohio	+0.3	+0.1	+3.1	+1.1
Louisville, Ky.-Ind.	+6.3	+6.8	+1.0	-1.1
Gary-Hammond-East Chicago, Ind.	-4.2	-6.3	+11.4	-5.0
Fort Worth, Texas	+6.2	+5.6	+12.0	+1.5
Nashville, Tenn.	+0.6	+2.0	-9.5	+3.1
Fort Lauderdale-Hollywood, Fla.	+0.1	+1.4	-7.0	+22.7
Orlando, Fla.	+5.1	+4.5	+8.7	+7.5
Charlotte, N. C.	-1.9	-0.3	-8.7	+4.4
Beaumont-Port Arthur, Texas	-5.7	-4.3	-11.8	-5.8
Chattanooga, Tenn.-Ga.	-4.1	-3.2	-10.2	-3.1
Shreveport, La.	-3.3	-3.2	-3.7	-6.0
Little Rock-North Little Rock, Ark.	+3.2	+6.1	-10.9	+5.4
Columbia, S. C.	+1.1	+6.0	-14.4	+2.2
Charleston, S. C.	-0.8	+1.9	-5.8	+4.4
Baton Rouge, La.	+16.0	+25.8	-19.0	+0.8
West Palm Beach, Fla.	+9.2	+10.2	+4.2	+15.7
Newport News-Hampton, Va.	+19.5	+17.6	+24.1	+8.9
Jackson, Miss.	+2.0	+6.7	-11.2	+2.5
Columbus, Ga.-Ala.	-1.6	+0.9	-8.3	+6.7
Augusta, Ga.-S. C.	+4.9	+6.7	-0.8	+0.9
Montgomery, Ala.	(Z)	+6.8	-18.8	-3.4
Savannah, Ga.	-11.7	-13.5	-7.4	-6.3
Macon, Ga.	-10.3	-5.7	-22.4	+2.1

1/ Net employed migrants as a percent of average base population, 1960-65. Includes military.
2/ Includes military. 3/ Includes nonwhite races other than Negro. (Z) Less than 0.05.

Appendix Table 1.--NET MIGRATION RATES, BY SEX, FROM CENSUS AND CWHS: 1955-60 AND 1960-65

Region, Division, and State	Male		Female	
	1955-60 Census 1/	1960-65 CWHS 2/	1955-60 Census 1/	1960-65 CWHS 2/
United States, Total	-	-	-	-
Regions				
Northeastern States	-2.0	-0.2	-1.4	-0.6
North Central States	-2.2	-1.5	-1.6	-1.6
The South	+0.2	-0.7	-0.3	-0.7
The West	+6.5	+4.4	+5.9	+5.5
Northeast				
New England	-0.6	-0.1	-0.8	-0.4
Middle Atlantic	-2.4	-0.2	-1.6	-0.7
North Central				
East North Central	-1.7	-1.4	-1.0	-1.1
West North Central	-3.3	-2.0	-3.2	-2.8
South				
South Atlantic	+2.7	-0.9	+1.9	-0.1
East South Central	-3.2	-1.5	-3.7	-1.8
West South Central	-1.3	+0.2	-1.4	-1.1
West				
Mountain	+4.2	+0.8	+4.7	+2.9
Pacific	+7.2	+5.4	+6.3	+6.2
New England				
Maine	-2.9	-5.6	-3.9	-5.1
New Hampshire	+1.7	-1.5	+1.1	+0.7
Vermont	-2.6	-1.5	-2.9	+1.0
Massachusetts	-1.5	-0.6	-1.2	+0.6
Rhode Island	-0.2	-1.4	-2.9	-1.0
Connecticut	+1.6	+3.2	+1.9	-1.1
Middle Atlantic				
New York	-3.1	-2.0	-2.1	-1.6
New Jersey	+2.3	+8.9	+2.2	+6.0
Pennsylvania	-3.8	-2.3	-2.7	-2.8
East North Central				
Ohio	-1.0	-2.9	-0.1	-1.5
Indiana	-1.2	-0.1	-1.1	(Z)
Illinois	-1.4	-0.1	-1.0	-0.9
Michigan	-3.2	-1.2	-2.0	-1.6
Wisconsin	-1.5	-2.9	-0.9	-1.3
West North Central				
Minnesota	-1.5	(Z)	-0.8	-2.1
Iowa	-4.8	-7.3	-4.3	-3.9
Missouri	-1.6	-1.5	-2.1	-3.7
North Dakota	-6.5	-4.3	-7.9	-11.2
South Dakota	-7.1	-7.3	-7.0	-4.4
Nebraska	-5.1	-4.9	-5.1	-2.0
Kansas	-4.1	+3.4	-4.1	+1.6

Appendix Table 1 continued

Region, Division, and State	Male		Female	
	1955-60 Census 1/	1960-65 CWHS 2/	1955-60 Census 1/	1960-65 CWHS 2/
South Atlantic				
Delaware	+4.6	+12.2	+4.9	+3.7
Maryland	+3.7	+7.2	+3.5	+7.3
District of Columbia	-10.8	-22.3	-9.6	-15.4
Virginia	+2.6	+1.9	+1.0	+3.4
West Virginia	-10.1	-5.9	-9.6	-4.8
North Carolina	-1.4	-2.2	-2.8	-2.5
South Carolina	-0.5	+0.6	-3.8	-0.2
Georgia	-0.7	-4.3	-1.7	-1.5
Florida	+16.8	(Z)	+17.6	+1.1
East South Central				
Kentucky	-4.3	-0.8	-5.1	-0.5
Tennessee	-2.7	+0.9	-3.0	-2.6
Alabama	-2.2	-3.6	-2.4	-1.1
Mississippi	-4.0	-3.9	-4.8	-2.6
West South Central				
Arkansas	-5.5	+1.0	-5.3	+1.0
Louisiana	-0.5	+4.2	-0.2	-4.2
Oklahoma	-4.0	-1.6	-4.4	-1.6
Texas	-0.1	-0.7	-0.3	-0.6
Mountain				
Montana	-3.1	-3.9	-3.2	-8.6
Idaho	-1.9	-7.8	-2.9	-5.6
Wyoming	-2.5	-11.0	-1.8	-3.6
Colorado	+3.8	+3.1	+4.5	+4.2
New Mexico	+6.0	-13.9	+5.5	-3.3
Arizona	+13.5	+6.8	+14.7	+10.5
Utah	+0.7	-0.4	+1.0	+0.2
Nevada	+8.3	+29.0	+10.1	+28.1
Pacific				
Washington	+1.6	+1.7	+1.0	+0.2
Oregon	-0.9	+0.4	-0.5	+0.5
California	+9.1	+6.8	+8.3	+7.9
Alaska	+10.1	+19.3	+5.5	+18.3
Hawaii	+5.0	-3.3	-0.2	+3.5

1/ Net migrants as a percent of census population 15-64 years of age. Includes military.

2/ Net employed civilian migrants as a percent of average base population in the Work History File, 1960-65. Includes an unknown number of immigrants from abroad.

(Z) Less than 0.05.

Appendix Table 2.--PERCENT NEGRO OF TOTAL POPULATION 15-64 YEARS OF AGE AND OF TOTAL WORKERS IN CWHS FILE FOR SMSA'S WITH 50,000 OR MORE NEGRO POPULATION IN 1960

Standard Metropolitan Statistical Area 1/	Percent Negro of Total Population 15-64 in 1960	Percent Negro of Total Workers in CWHS File, 1960
250,000+ Negro population, 1960		
New York, N. Y.	11.5	11.5
Los Angeles-Long Beach, Calif.	7.5	6.6
Chicago, Ill.	13.6	11.8
Philadelphia, Pa.-N. J.	15.1	13.8
Detroit, Mich.	14.5	11.4
St. Louis, Mo.-Ill.	13.2	11.1
Washington, D. C.-Md.-Va.	23.3	24.6
Cleveland, Ohio	13.3	11.9
Baltimore, Md.	20.3	19.7
Houston, Texas	18.6	16.8
New Orleans, La.	28.1	26.2
Memphis, Tenn.-Ark.	34.0	27.8
100,000-249,000 Negro population, 1960		
San Francisco-Oakland, Calif.	8.1	5.8
Pittsburgh, Pa.	6.5	5.4
Newark, N. J.	12.9	10.9
Cincinnati, Ohio-Ky.-Ind.	10.2	8.4
Dallas, Texas	13.7	12.1
Kansas City, Mo.-Kans.	10.2	9.1
Atlanta, Ga.	21.5	18.1
Indianapolis, Ind.	10.5	8.5
Miami, Fla.	13.9	15.9
Birmingham, Ala.	29.5	25.0
Norfolk-Portsmouth, Va.	23.9	31.1
Greensboro-Winston-Salem-High Point, N. C.	18.8	16.4
Jacksonville, Fla.	22.0	22.7
Richmond, Va.	25.0	23.4
Mobile, Ala.	28.4	25.2
50,000-99,000 Negro population, 1960		
Boston, Mass.	2.9	2.7
Buffalo, N. Y.	6.2	5.9
Milwaukee, Wisc.	4.5	4.5
Tampa-St. Petersburg, Fla.	11.6	11.8
Columbus, Ohio	10.5	8.2
Dayton, Ohio	9.5	7.0
Louisville, Ky.-Ind.	10.9	10.3
Gary-Hammond-East Chicago, Ind.	14.2	12.1
Fort Worth, Texas	9.9	9.0
Nashville, Tenn.	17.9	13.0
Fort Lauderdale-Hollywood, Fla.	15.0	17.6
Orlando, Fla.	16.2	14.2
Charlotte, N. C.	22.2	20.4
Beaumont-Port Arthur-Orange, Texas	19.3	17.9
Chattanooga, Tenn.-Ga.	16.4	13.6
Shreveport, La.	30.4	28.7
Little Rock-North Little Rock, Ark.	17.7	16.9
Columbia, S. C.	25.9	26.4
Charleston, S. C.	33.6	35.6
Baton Rouge, La.	29.8	25.6
West Palm Beach, Fla.	23.2	19.2
Newport News-Hampton, Va.	26.4	28.5
Jackson, Miss.	36.3	25.9
Columbus, Ga.-Ala.	26.8	27.0
Augusta, Ga.-S. C.	26.3	24.1
Montgomery, Ala.	34.1	27.0
Savannah, Ga.	32.6	31.6
Macon, Ga.	28.3	27.8

1/ As defined in 1968.

Appendix Table 3.--STANDARD ERRORS OF 1960-65 NET MIGRATION RATES FOR STATES BASED ON SOCIAL SECURITY DATA

(1 Sigma)

Region, Division, and State	All Classes	White 1/	Negro
United States, Total			
Regions			
Northeastern States	0.2	0.2	0.6
North Central States	0.2	0.2	0.6
The South	0.2	0.2	0.5
The West	0.3	0.3	1.4
Northeast			
New England	0.4	0.4	2.7
Middle Atlantic	0.2	0.2	0.6
North Central			
East North Central	0.2	0.2	0.7
West North Central	0.4	0.4	1.9
South			
South Atlantic	0.3	0.3	0.6
East South Central	0.5	0.5	1.0
West South Central	0.4	0.4	0.8
West			
Mountain	0.7	0.8	5.9
Pacific	0.3	0.3	1.3
New England			
Maine	1.4	1.4	(B)
New Hampshire	1.8	1.8	(B)
Vermont	2.6	2.6	(B)
Massachusetts	0.5	0.5	4.1
Rhode Island	1.3	1.3	(B)
Connecticut	0.7	0.8	3.8
Middle Atlantic			
New York	0.2	0.3	0.8
New Jersey	0.4	0.6	1.8
Pennsylvania	0.3	0.3	1.2
East North Central			
Ohio	0.4	0.4	1.3
Indiana	0.6	0.6	2.5
Illinois	0.4	0.4	1.1
Michigan	0.4	0.4	1.2
Wisconsin	0.6	0.6	(B)
West North Central			
Minnesota	0.7	0.7	(B)
Iowa	0.9	0.9	(B)
Missouri	0.7	0.7	2.0
North Dakota	2.6	2.6	(B)
South Dakota	2.3	2.3	(B)
Nebraska	1.3	1.3	(B)
Kansas	1.2	1.2	6.3

Appendix Table 3 continued

Region, Division, and State	All Classes	White 1/	Negro
South Atlantic			
Delaware	2.4	2.6	6.2
Maryland	0.8	1.0	1.8
District of Columbia	2.0	2.4	3.4
Virginia	0.8	0.9	1.6
West Virginia	1.2	1.2	5.1
North Carolina	0.6	0.7	1.3
South Carolina	1.0	1.1	1.9
Georgia	0.8	0.9	1.4
Florida	0.8	0.9	1.6
East South Central			
Kentucky	1.0	1.1	3.0
Tennessee	0.8	0.9	1.8
Alabama	0.9	1.1	1.7
Mississippi	1.2	1.5	2.0
West South Central			
Arkansas	1.4	1.6	2.8
Louisiana	0.9	1.2	1.5
Oklahoma	1.1	1.2	4.7
Texas	0.4	0.5	1.0
Mountain			
Montana	2.2	2.2	(B)
Idaho	2.6	2.6	(B)
Wyoming	3.6	3.6	(B)
Colorado	1.4	1.4	(B)
New Mexico	2.4	2.4	(B)
Arizona	1.7	1.7	(B)
Utah	1.7	1.7	(B)
Nevada	3.9	4.0	(B)
Pacific			
Washington	0.9	0.9	(B)
Oregon	1.2	1.2	(B)
California	0.3	0.3	1.4
Alaska	3.7	3.7	(B)
Hawaii	2.0	2.0	(B)

(B) Having less than 50,000 Negro population 15-64 years of age in 1960.

1/ Includes nonwhite races other than Negro.

Appendix Table 4.--STANDARD ERRORS OF 1960-65 NET MIGRATION RATES FOR SMSA'S BASED ON SOCIAL SECURITY DATA

(1 Sigma)

Standard Metropolitan Statistical Area 1/ Rank and Size Group	All Classes	White 2/	Negro
250,000+ Negro population, 1960			
New York, N. Y.	0.4	0.4	1.1
Los Angeles-Long Beach, Calif.	0.7	0.7	2.5
Chicago, Ill.	0.5	0.6	1.3
Philadelphia, Pa.-N. J.	0.7	0.7	1.6
Detroit, Mich.	0.7	0.7	1.7
St. Louis, Mo.-Ill.	1.0	1.1	2.6
Washington, D. C.-Md.-Va.	1.4	1.7	2.4
Cleveland, Ohio	1.1	1.2	2.7
Baltimore, Md.	1.0	1.2	2.1
Houston, Texas	1.7	1.8	3.6
New Orleans, La.	2.0	2.5	3.3
Memphis, Tenn.-Ark.	2.3	2.8	3.5
100,000-249,000 Negro Population, 1960			
San Francisco-Oakland, Calif.	1.2	1.3	4.9
Pittsburgh, Pa.	0.9	1.0	3.6
Newark, N. J.	1.3	1.4	4.0
Cincinnati, Ohio-Ky-Ind.	1.4	1.4	4.2
Dallas, Texas	1.7	1.8	4.7
Kansas City, Mo.-Kans.	1.6	1.7	4.3
Atlanta, Ga.	1.7	1.9	3.5
Indianapolis, Ind.	1.6	1.7	4.7
Miami, Fla.	2.1	2.3	4.6
Birmingham, Ala.	2.1	2.5	3.9
Norfolk-Portsmouth, Va.	3.1	3.9	5.1
Greensboro-Winston-Salem-High Point, N. C.	2.1	2.3	4.8
Jacksonville, Fla.	3.5	4.0	7.4
Richmond, Va.	2.3	2.7	4.4
Mobile, Ala.	3.9	4.6	7.1
50,000-99,000 Negro population, 1960			
Boston, Mass.	0.8	0.8	5.3
Buffalo, N. Y.	1.2	1.2	4.5
Milwaukee, Wisc.	1.2	1.2	5.4
Tampa-St. Petersburg, Fla.	2.6	2.7	7.3
Columbus, Ohio	2.1	2.2	6.8
Dayton, Ohio	1.9	2.0	6.6
Louisville, Ky.-Ind.	1.8	1.9	5.3
Gary-Hammond-East Chicago, Ind.	2.0	2.2	4.3
Fort Worth, Texas	2.8	2.9	8.9
Nashville, Tenn.	2.5	2.7	6.4
Fort Lauderdale-Hollywood, Fla.	4.6	5.0	10.7
Orlando, Fla.	4.5	4.8	12.4
Charlotte, N. C.	3.0	3.4	6.3
Beaumont-Port Arthur, Texas	3.5	3.8	8.1
Chattanooga, Tenn.-Ga.	3.0	3.3	7.5
Shreveport, La.	4.6	5.6	7.3
Little Rock-North Little Rock, Ark.	3.6	4.1	7.2
Columbia, S. C.	3.8	4.4	7.2
Charleston, S. C.	4.4	5.8	6.7
Baton Rouge, La.	4.9	5.7	9.8
West Palm Beach, Fla.	5.0	5.5	11.7
Newport News-Hampton, Va.	4.7	5.6	8.6
Jackson, Miss.	3.9	4.7	6.7
Columbus, Ga.-Ala.	4.7	5.7	7.6
Augusta, Ga.-S.C.	4.2	5.0	7.4
Montgomery, Ala.	4.6	5.6	7.9
Savannah, Ga.	4.4	5.5	7.4
Macon, Ga.	5.2	6.5	8.6

1/ As defined in 1968. 2/ Includes nonwhite races other than Negro.

MOBILITY ANALYSIS - AN APPROPRIATE TECHNIQUE IN SMALL AREA DEMOGRAPHY

Arnold R. Post, Fels Center, University of Pennsylvania

I wish to acknowledge that this paper derives from conversations with one Robert Reams. He was a supervisor of special censuses for the Census Bureau in the late 1950's and had developed a rule of thumb for estimating current population which he found to be quite accurate. In a variety of areas, 1958 populations were approximately equal to 90 percent of their 1950 populations plus about 3.75 times their increase in dwelling units since 1950.

Actually, it may be observed that populations behave with considerable regularity for given changes in their numbers of households (or housing supply). Multiple linear regressions for collections of areas in Pennsylvania indicate that their censused populations may be reliably estimated as deriving from the populations previously censused and subsequent changes in number of households. More than 99 percent of the variance in final population is 'explained,' and the coefficients of the two independent variables are both very highly significant. Standard errors and constant terms tend to be relatively small, and residuals appear to be reasonably well distributed according to normal expectations.

Several analyses have been done on the following sets of areas:

A. A collection of counties in metropolitan areas at least partly in Pennsylvania and having at least 3,000 nonwhite residents, but including only those counties or parts of Philadelphia with at least 97 percent of population in households (28 cases);

B. A collection of 20 randomly selected Pennsylvania counties; and

C. The collection of 427 census tracts (as of 1950) in the Philadelphia Standard Metropolitan Statistical Area (SMSA), but outside the city proper.

Some of the results are shown in Table 1, where the figures in parentheses show the t-values of the coefficients of the estimating equations. A t-value greater than 3.0 usually fosters belief that a real-world dependency has not been disproved by the data at hand. As can be seen, these t-values range from 5.57 to 205.3.

The random sample of counties and the suburban census tracts contained predominantly white populations resident mostly in single family dwellings. Among the counties were several which lost population during the 1950's, while the tracts were characterized mostly by rapid growth. The metropolitan counties are associated with higher percentages of nonwhite population and apartment dwellers.

Relations 3,5 appear to be distinct from each other, perhaps reflecting the different types of areas they have been drawn from. Both, however, can be regarded as within the range of random deviation about relation 4, which was derived from the probability sample.

The coefficients of relation 4 have reasonable values. The coefficient of initial population is 89.8 percent, which is not far off the tenth power of .99. Janet Abu-Lughod and Mary Mix Foley

Table 1
Linear Regression Equation Parameters For Estimating Population After Ten Years, Given Initial Population and Subsequent Change in Households
(t-values in parenthesis)

Data Set	Decade	Initial Population	Change in Households	Constant Term	Mean Value Est. Pop. ($\bar{Y}$)	Standard Error No.	Standard Error % of $\bar{Y}$	Relation Number
A	1930-40	.9053 (41.4)	3.3425 (5.57)	3,753	221,006	8,803	4.0 %	(1)
	1940-50	.9116 (126.4)	3.6951 (25.1)	1,398	240,652	4,295	1.8 %	(2)
	1950-60	.9234 (80.3)	3.8495 (20.0)	516	267,822	12,612	4.7 %	(3)
B	1950-60	.8982 (53.6)	3.9583 (20.2)	2,065	86,736	2,410[1]	2.8 %	(4)
C	1950-60	.8858 (145.9)	4.1010 (205.3)	28	5,293	343[2]	6.5 %	(5)

(1) 65% of residuals less than one standard error, 10 positive (max. 5,425), 10 negative (min. -3,341)
(2) 95% of residuals less than one standard error, about 60% positive.

Source: U.S. Censuses of Population for 1930,1940,1950, and 1960. Private families, occupied dwelling units, and occupied housing units have been treated as if synonymous.

have reported (1) an estimate to the effect that in 1955, about 1 percent of the population engaged in household formation. Presumably, then, 99 percent did not; and over the course of the decade, assuming constancy, about 89 percent would be resident in the same number of households as at the beginning of the decade. The coefficient of household change (3.96) lies near the mid-point of a range defined by the average size of all husband-wife households (3.7 persons) and the average size of such households with husband under 45 years of age (4.3 persons.) (2) The coefficient is therefore consistent with the notion that gains in households are principally the work of young people, as has been well established by Ned Shilling. (3)

These relations in Table 1 were derived from the period 1930-60. Considering that this period spans economic extremes of depression and boom, the similarities among them are notable. The coefficients of initial population are less than 1.0 and those of household change are not unwieldy as average household sizes. Both have proper sign.

I can't, at this time, present statistical findings for the nation's large metropolitan areas, but I have drawn some pictures to show population contours for some of them. For the very largest metropolitan areas (1950 SMA's, generally) the populations of 1950 are shown in Figure 1 as a joint function of 1940 population and net gain in households. Similar information is shown in Figure 2 for 1960 populations of SMA areas of between 200-700 thousand persons as of 1950 and with household gains of less than 105,000. The data have been scaled assuming 4 persons per marginal household. Under the assumed hypothesis, the contours should be straight, equally spaced, and parallel with a slope of -45 degrees. In Figure 1, the contours can be seen to approach the vertical axis at values greater than their own, as should be the case.

Now, given such a well-defined and appropriate collection of regression equations which seem to be broadly applicable, it is only reasonable to suppose that they index a process. Since the coefficient of initial population presumably indicates a fraction of population apt to remain in a stable community (one characterized by no net change in number of households), it can be termed a persistence rate. The coefficient of household change can be considered a rate of marginal population change since it would indicate the average size of marginal households. Accordingly, we would have two concepts, persistence and marginal change, which are somewhat analagous to the traditional concepts of natural increase and net migration, except that we are here dealing with a "natural increase in housing need" and the mobility occasioned to satisfy it. More concretely, a population is assumed to generate additional family or household heads in the course of a decade, and they will export themselves to areas where the desired quarters are available, if they can manage it, whether across the street or across the nation. They will also take with them wives and potential children, among others, as they see fit.

Commonly accepted demographic algorisms rely on concepts of natural increase (surplus of births over deaths) and net migration; and it has been asserted many times that population change in a given place derives either from natural increase or net migration. This assertion carries with it an implication that net-migrants are both sterile and immortal. Indeed, the implication appears explicitly in some matrix models of population growth. As a migrant, myself, currently, I take exception.

Traditional application of these concepts is often called "cohort survival analysis." The initial population of a given place in this context, provides a source of people likely to bear and survive anywhere. The concept of net migration re-introduces boundaries very much as an afterthought. Migrants are, after all, only ordinary mortals who have been reclassified. Some attrition among cohort survivors has to be allowed for somewhere in any realistic accounting of the two classes of people, that is, today's migrants are yesterday's naturalized residents of some other place.

Algorisms relying on the concepts of persistence and marginal change have been categorized as dealing in "mobility analysis." These concepts relate nicely to population trends as defined in geographic partitions. For each area, conditional allowance is made both for departures (as a function of initial population) and arrivals or additional departures (as a function of household change.) The initial population in this context provides a source of people apt to survive and remain (on net) in a given place; and the additional households, if any, provide accommodation for new families or families moving about. To my mind, mobility analysis is richer in spatial context and logically more of a piece than cohort survival analysis and has much to recommend it. Statistically significant results have also been obtained from age-specific data.

Mobility analysis, however, is not any easier to apply than cohort survival analysis; and it does require independent estimates of postcensal or future changes in households to explore likely patterns of population development.

For current estimates of household change, my preference is to rely on school enrollment data, interpreting them with the aid of census cross-tabulations of 5 to 19 year-olds by grade of school in which enrolled, using three-year grade groups and five-year age groups. With the aid of the cross-tabs, one can develop one operator to generate an age distribution from enrollment data and another to generate an enrollment distribution from age data. (The cross-tabs are published by single years of age and grade among the detailed statistics for each state.)

Current school enrollment data will not automatically be available for each municipality since public school districts sometimes don't coincide with municipal boundaries and private school service areas may be quite independent of them. The data, however, are worth scratching for, it being important to secure the equivalent of pupil counts as of April.

Figure 1

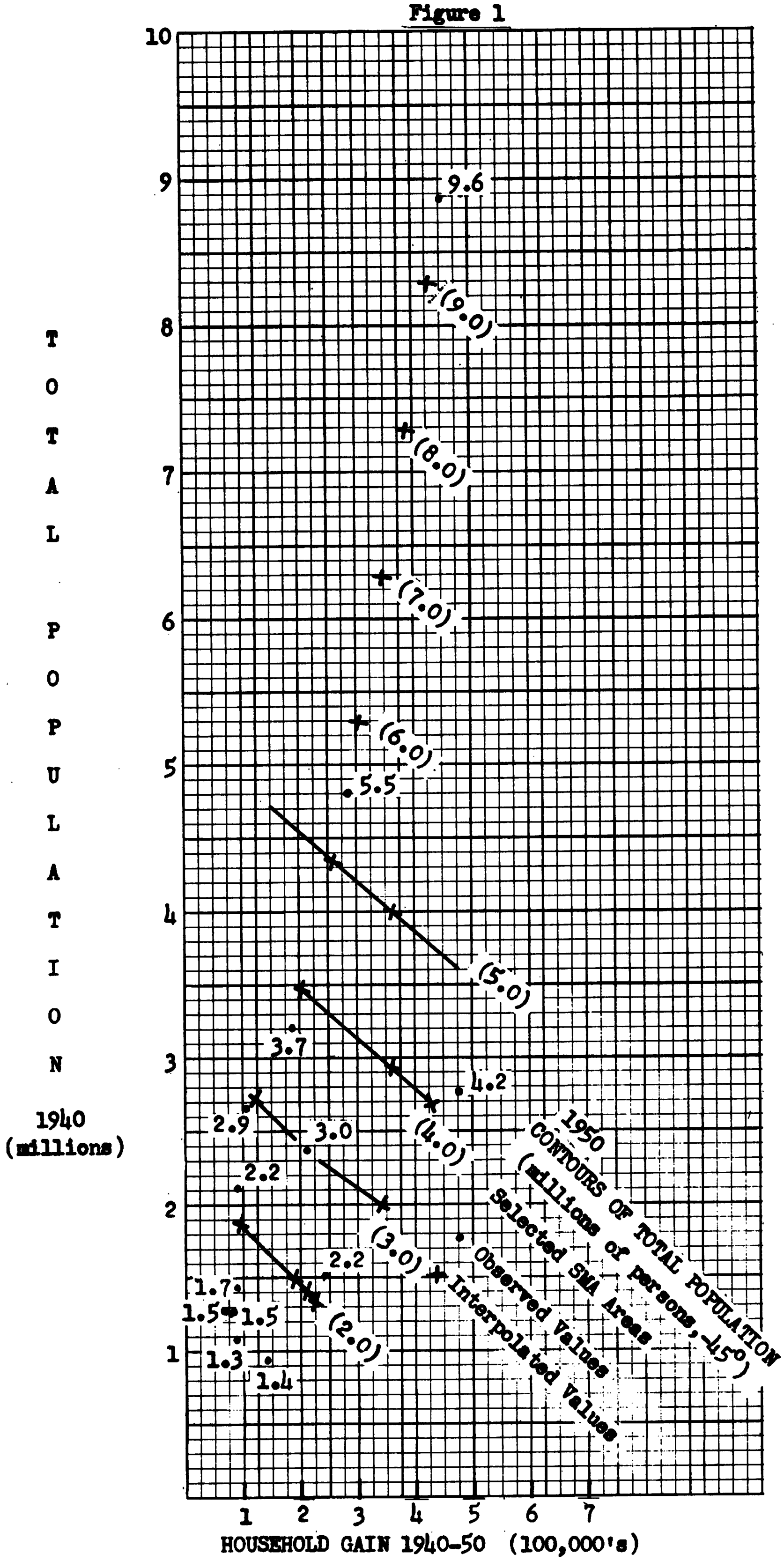
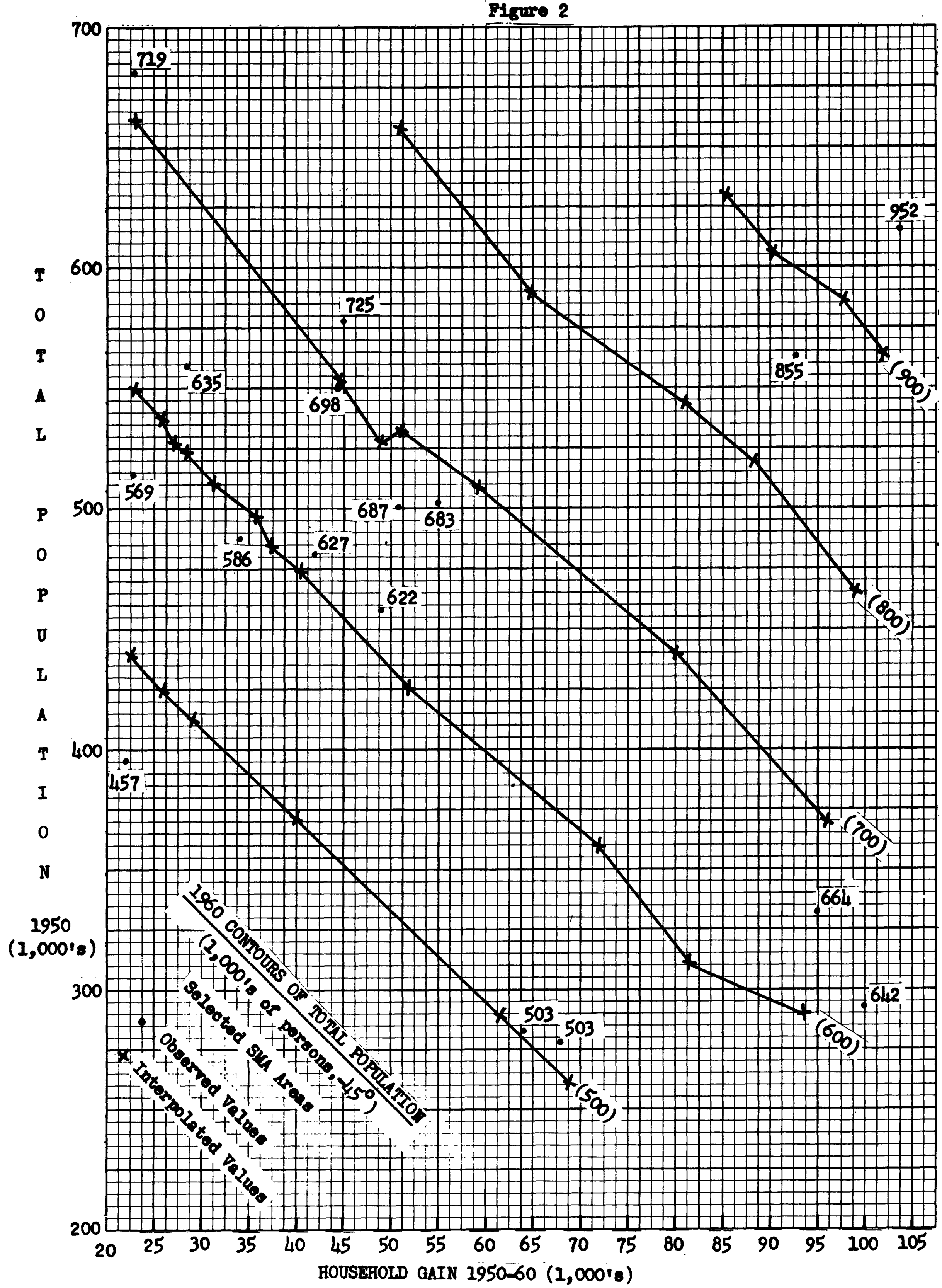

TOTAL POPULATION
1940
(millions)
HOUSEHOLD GAIN 1940-50 (100,000's)
1950
CONTOURS OF TOTAL POPULATION
(millions of persons, -45°)
Selected SMA Areas
Observed Values
Interpolated Values
(2.0)
(3.0)
(4.0)
(5.0)
(6.0)
(7.0)
(8.0)
(9.0)
9.6
5.5
4.2
3.7
3.0
2.9
2.2
2.2
1.7
1.5
1.5
1.3
1.4

Figure 2


TOTAL POPULATION 1950 (1,000's)
HOUSEHOLD GAIN 1950-60 (1,000's)
1960 CONTOURS OF TOTAL POPULATION
(1,000's of persons, 45°)
Selected SMA Areas
Observed Values
Interpolated Values
(500)
(600)
(700)
(800)
(900)
719
725
635
698
569
687
683
586
627
622
457
855
952
664
642
503
503
700
600
500
400
300
200
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
105

With enrollment data at hand, one attacks them from two directions. (1) On the basis of the last census, a persistent population of school-age is estimated as if there had been no change in the number of households. An average number of these children per household is also calculated. Then a disaggregated estimate of their enrollment by age and grade group is prepared to provide an index of the age distribution in the grade groups currently. (2) The actual enrollments by grade group are assumed to have the same relative age distribution as enrollments from the persistent population. The age groups are then summed and blown up to indicate the age distribution which generated the actual enrollments. This expansion is required since not all school-age children will actually be in school.

The difference between the presumed actual and the persistent populations (5-19) can then yield an estimate of net household change according to which the total population and its age distribution can be estimated.

At this point, after the direction of net migration can be determined, useful reference to vital statistics can be made. The argument, however, has to be interpretive. Abu-Lughod and Foley in the work cited, indicate that changes in household size appear to be a main factor associated with a fimily's taking new quarters. Birth statistics, then, would overstate preschoolers for a population supporting out-migration and understate them for a population augmented by in-migration, i.e., babies seem to be a cause of mobility and they travel with their parents.

The above paragraphs sketch some of the procedures available to make mobility analysis operational, and difficulties have scarcely been mentioned. The real world and the future continue to present problems that no amount of arithmetic can render precise. I think it is fair to say that the past is about as indefinite as the future is improbable, and the two are related by an uncertain, though actual, present. In the late 1950's, I, for one, did not anticipate the development of near revolutionary conditions in many of our major cities; and some of my estimates are doubtless less apposite for that. In terms of the present discussion, I would expect these conditions to affect rates of persistence and marginal change by influencing especially the mobility of mature families. Although I did anticipate the declining trend in births in fair degree, I didn't foresee the growth of the public share of total enrollments.

Accepting such matters as part of the rules of the game, we can still suppose some regularities exist. Glancing again at Table 1, it may be noted that between 1930 and 1960 (relations 1,2,3), rates of persistence and marginal change seem to be directly related with both rates rising. My own hunch is that some explanation is offered by the concurrently rising trend in large area rates of natural increase. What with reduced rates of natural increase now prevailing, it may be that current rates of persistence and marginal change are also now at lower levels. It may also be noted, however, that as regards relations 3,5, their coefficients are inversely related with the greater spread being characteristic of the richer populations. I incline to consider this as evidence that richer populations can form new households at greater convenience. Although I can't quantify these matters, I anticipate that the greater mobility of mature families, if any, will tend to make both rates less well defined in the present decade, that with lowered natural increase, the rates will define a somewhate shorter vector, and that with continued prosperity, the spread between them will tend to remain wide.

As for internal matters, Table 2 exhibits a good deal of variety among ten-year rates of persistence and marginal change as derived from analysis of age-specific data for the metropolitan counties. There are very low rates of marginal change for 15-29 year olds in 1960; and their level may relate to the scarcity of persons born between 1930 and 1944 in the total population, suggesting a future avenue of research towards a lagged model. Ten-year age-specific rates were also derived from the random sample of 20 Pennsylvania counties; and these rates may be more appropriate in suburban and rural areas. They are shown in Table 3.

Since rates of community development may change from one pentad to the next (I don't like to say quinquennium, except to refuse to say it), it is desirable to work in terms of five-year intervals. Conjectures, however, as to the values of five-year rates are simply conjectures at this time. A special census was done of the Cleveland SMSA in 1965, and these data may provide a springboard.

A comparison of persistence and survival rates is given in an article scheduled to appear in the November (1969) issue of the Journal of the American Institute of Planners. A comparison is also offered there of age distributions of marginal changes and net migratory increments. In addition, it is noted that relation 5, Table 1, derived from census tract data, yields good results with respect to the population of the whole Philadelphia SMSA and to that of each of its constituent counties, including Philadelphia, itself, whose census tracts did not contribute to the estimating equation. The bias is on the order of -1.5 percent. (Note: Some may wish to investigate these relationships in their own regions of interest. If so, it is necessary to examine an area large enough to provide residence for by far the greater part of both its initial and final populations. All the census tracts in one suburban county may not be sufficient if the county, say, has doubled its population in the period under review. In such a case, the distribution of initial population by tract is not apt to be significant.)

Mobility analysis has been employed as the basis for a computerized routine to yield estimates of the short run future for a variety of suburban and rural school districts. Analysis done by hand for Philadelphia indicates (1) that the city, with a gain of only 20,000 households, is continuing to lose population. A population of about 1.9 million is anticipated for 1970, about 5 percent less than 1960's

Table 2
Ten Year Age-Specific Rates, Metropolitan Counties

Terminal Age Group	Persistence 1930-40	Persistence 1940-50	Persistence 1950-60	Marginal Change 1930-40	Marginal Change 1940-50	Marginal Change 1950-60
0-4	.1053[1]	.1510[1]	.2047[1]	.3695	.5714	.5323
5-9	.1117[2]	.1260[2]	.1882[2]	.4020	.4193	.5176
10-14	.9389	.9016	.8187	.1393	.2149	.3457
15-19	.8805	.8373	.8054	.2439	.1883	.1812
20-24	.8874	.8265	.7662	.2454	.2390	.1128
25-29	.8397	.7766	.8006	.3216	.4169	.2406

(1) Generation rates based on initial population 5-39.
(2) Generation rates based on initial population 10-44.
Source: See note, Table 1.

Table 3
Ten Year Age-Specific Rates,
20 Pennsylvania Counties, 1950-60

Terminal Age Group	Persistence	Marginal Change
0-4	.091[1]	.665
5-9	.096[1]	.560
10-14	.818	.356
15-19	.751	.307
20-24	.575	.359
25-29	.635	.376

(1) Generation rates based on total initial population.
Source: See note, Table 1.

2.0 million; (2) that the proportion of population Negro appears to be increasing more rapidly than during the 1950's and will reach 37 percent by 1970; and (3) that the public share of a nearly constant total enrollment seems apt to increase by about 7 percent on a base of 410,000 by 1980.

The Philadelphia estimates do not represent a consensus. Other analysts see a small gain for the city population, slower growth of the proportion Negro, and less than a 28,000 gain in public school enrollments in the coming decade. The 1970 census should lay some doubts to rest. Although my goal as regards accuracy is to be most accurate among those who overestimate, I should think that a census of 1.95 million or less total population would indicate that some skill has been achieved in the application of mobility analysis.

Overall, the regularities that have been observed do encourage me to believe that locally standard patterns of population growth and development have become a little more apparent.

REFERENCES

(1) Janet Abu-Lughod and Mary Mix Foley, "Consumer Strategies," Part II of Nelson Foote, et al., Housing Choices and Housing Constraints (New York: McGraw-Hill, 1960), Table 19, p.99 and p.100 (footnote).

(2) U.S., Bureau of the Census, Current Population Report, P-20, No. 173 (June 25,1968), Tables 1 and 5.

(3) Ned Shilling, "Net Household Formation - A Demographic Analysis," (unpublished Master's Essay, Columbia University, 1955), cited by Louis Winnick, American Housing and Its Use (New York: Wiley and Sons, 1957), p.81.

INSTANT SUPERMATRIX SETUPS FOR MULTIVARIATE, ITERATIVE, AND MULTI-EQUATION MODELS

Irving H. Siegel, W.E. Upjohn Institute for Employment Research*

Aim and Background

My purpose is to illustrate the usability of data supermatrices in more complex cases of least-squares estimation than I have been treating earlier. My motivation derives in part from inquiries occasioned by a paper that I presented at the August 1968 meeting of the American Statistical Association [1] and by another that appeared in the December 1968 American Statistician [2]. An additional stimulus has been provided by the development of a new computer program for regression analysis that takes account of the special features of least-squares supermatrices. The availability of such a program, discussed in another paper presented at this meeting, should encourage the practical utilization of the supermatrix approach.

In the classical least-squares cases, which involve unweighted or weighted observations (with no auxiliary conditions on the unknown parameters), the supermatrix method permits instant organization of the data, without any prior processing at all, into the algebraic equivalents of normal equations. If, in the weighted case, some prior arithmetic processing is allowed, an alternative supermatrix system may be written that is more compact but still algebraically equivalent. As a rule, there is a tradeoff between the acceptable amount of prior computation and the acceptable size of the supermatrix system.

It was for the relatively simple classical cases that the supermatrix approach was originally developed. In my earliest gropings toward this method (see my 1941 papers in the Journal of the American Statistical Association [3], [4]), I used a weighted least-squares model and rectangular data matrices in studying the relationship between aggregative index numbers that embody different sets of weights.

I have already demonstrated the extensibility of supermatrix design beyond the simplest and most familiar situations. Thus, all or most prior arithmetic operations are avoidable (a) in the fitting of a straight line when all observations are subject to error (othogonal regression represents a special subcase)[2] and (b) in least-squares adjustment when the unknown constants are subject to side constraints [5].

Now, I consider additional, more difficult, instances. If these instances may be said to exemplify a common theme, it is "multiplicity." Thus, I say something below about the estimation of parameters for a function with multiple variables subject to error. I also offer two illustrations of multiple-step approximation; one is iterative in the sense of ultimate approach to least-squares estimation, and the other is iterative in the sense that the original linear model undergoes refinement by the addition of parameters and observations. Finally, I make some remarks on models involving multiple equations subject to disturbance.

Before proceeding with the discussion of new cases, I want to comment further on least-squares supermatrix systems. All of these have the general form $Fg = h$. Here, F is a square design matrix -- really, a supermatrix -- containing unprocessed observations on one or more regressors and providing coefficients for individual error terms. These coefficients state assumptions regarding the combination of errors. The design matrix may also include, if the problem requires, assumptions relating to additional Lagrangian terms. The unknown constants, the (usually) unknown observational errors (really, "residuals"), and the unknown Lagrangian multipliers (if required) are incorporated in a supervector, g. Another supervector, h, (usually) shows the observations on the dependent variable and the aggregate or summary conditions imposed on the errors and on the Lagrangian terms (the aggregates typically are zero).

F, g, and h have well-defined structures. That is, they are partitionable in rather obvious ways. The elements they contain are organized in characteristic "packages," according to the nature of the adjustment problem to be solved. Since these packages,

*The author's views should not be ascribed to the Upjohn Institute.

moreover, enter discernibly into simultaneous submatrix equations, the supermatrix system may be conceived as the result of a "stacking" procedure [5].

One of the ways of exhibiting the structure of the supermatrix system is to cast it into this form:

$$\begin{bmatrix} 0 & Q \\ P & I \end{bmatrix} \cdot \begin{bmatrix} d \\ r \end{bmatrix} = \begin{bmatrix} 0 \\ Y \end{bmatrix}$$

$$(F \cdot g = h) .$$

Here, 0 is a block of zeros; I stands for one or more (diagonal) identity matrices; P shows the observed values of the independent variables or regressors; Q gives the coefficients of the individual error terms (residuals); d represents a subvector of unknown parameters; r is the subvector of unknown errors; and Y typically refers to observed (sometimes, assumed or computed) values of the dependent variable. When the observations on the dependent and independent variables are subject to error, the elements of r may be composite. In general, $Qr = 0$ amounts to a simple direct statement of the normal equations in terms of the errors (i.e., residuals); and $Pd + Ir = Y$ (strictly, an identity rather than an equation) sets out the observed data in the form of the function, the prototype relationships, that is to be estimated.

In my first paragraph, I mentioned that another paper being presented at this ASA meeting deals with the programming of supermatrix systems (in FORTRAN) for efficient computer solution. My collaborator is Mr. Mac Shaibe, of the U.S. Bureau of Labor Statistics. As a result of his efforts, I expect supermatrix regression, which is not just a curious technical toy, to become more widely recognized as a potentially useful tool.

Multiple Variables Subject to Error

The multivariate case in which all observations are subject to error may be handled in (a) exactly the same way that I have treated the straight line with x's and y's subject to error [2] or (b) an alternative way that leads to a still larger supermatrix system. This larger system can be set up instantly with virtually no prior arithmetic processing. Some awkwardness may arise in actual computation, however, since the design matrix itself includes unknown parameters!

To work out the required supermatrix pattern, I have made use of what I call "normal identities" [6]. These identities, derived from the observation equations in a simple systematic manner, are reducible to normal equations when the appropriate error-affected aggregates are set equal to zero. Since not all of the error-affected aggregates in the identities can reasonably be assumed to vanish, however, we are left with more unknowns than equations and accordingly have to make additional assumptions to connect the unannihilated aggregates. In any event, the suppression and interconnection of aggregates are not matters of arbitrary choice; for the object is to obtain by the supermatrix method the same result that is given by the relevant normal equations. From a scrutiny of normal identities, we learn how to treat error aggregates for the derivation of the correct normal equations; and these normal equations have to be examined in turn for clues to their "explosion" into supermatrix equivalents.

For simplicity of exposition, I consider a function of the form $y = a + bx + cz$, where a, b, and c are the unknown parameters and where the observed values of y, x, and z are subject to error. Thus, I really start with equations (identities) of the form

$$y_i + s_i = a + bx_i + bt_i + cz_i + cu_i$$

$(i = 1, \ldots, n)$, where s_i, bt_i and cu_i represent unknown error terms that are to be "purged" from the observations. I call attention to the appearance of b and c as separate unknowns in the subvector d and also in combination with the error terms t_i and u_i, respectively, in the subvector r.

In addition to a block of zeros $(3n + 3, 3)$ in the "northwest" corner of the design matrix, I obtain these packages of elements for the supermatrix system:

$$Q = \begin{bmatrix} 1 \ldots & 1 \ldots & 1 \ldots \\ y_1 \ldots & x_1 \ldots & z_1 \ldots \\ z_1 \ldots & y_1 \ldots & x_1 \ldots \\ kx_1 & & by_1 \\ \ddots & mz_1 & \ddots \quad cy_1 \\ & \ddots & \ddots \end{bmatrix},$$

$$[P \; I] = \begin{bmatrix} 1 & x_i & z_i & 1 & 1 & 1 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots \end{bmatrix},$$

$$\begin{bmatrix} d \\ r \end{bmatrix} = [a \; b \; c \quad bt_1 \ldots \quad cu_1 \ldots \quad -s_1 \ldots]', \text{ and}$$

$$\begin{bmatrix} O \\ Y \end{bmatrix} = [0 \; 0 \; 0 \quad 0 \ldots \quad 0 \ldots \quad y_1 \ldots]'.$$

Dots indicate an extension to n terms. The symbols k and m appearing in Q represent constants to which values may be assigned according to assumptions desired with regard to relative variances. Thus, having only four normal equations but three parameters, a, b, and c, and three unknown and nonvanishing error aggregates, Σsy, $b\Sigma tx$, and $c\Sigma uz$, I interconnect the sums (which really represent variances) in this way: $\Sigma sy = k\Sigma tx = m\Sigma uz$. Notice that the terms kx_i and mz_i are the only ones in the design supermatrix inviting or requiring any prior computation.

<u>Two Iterative Cases</u>

In my first example of sequential approximation, I want to estimate the parameters of a straight line, but I choose to guess initially at the least-squares y's corresponding to the given x's. I could also introduce a second or third guess if I choose to do so; the supermatrix system is not overtaxed thereby. What is the supermatrix setup that assures "closure," that makes up for my bad guesses and yields the correct least-squares estimates of the parameters anyway?

For simplicity, I assume only one guessing round. I set up the supermatrix system for the observations $y_i = a + bx_i + e_i$ ($i = 1, \ldots, n$) in this manner:

$$Q = \begin{bmatrix} 1 & & 1 & & 1 & \ldots \\ & 1 & & 1 & & \\ x_1 & & x_2 & & x_3 & \ldots \\ & x_1 & & x_2 & & \ldots \end{bmatrix},$$

$$[P \; I] = \begin{bmatrix} 1 & & x_1 & & 1 & & & & \\ & 1 & & x_1 & & 1 & & & \\ 1 & & x_2 & & & & 1 & & \\ & 1 & & x_2 & & & & 1 & \\ 1 & & x_3 & & & & & & 1 \\ \vdots & \vdots & \vdots & \vdots & & & & & \ddots \end{bmatrix},$$

$$\begin{bmatrix} d \\ r \end{bmatrix} = [a'a'' \; b'b'' \quad e_1{}' e_1{}'' e_2{}' e_2{}'' \ldots]',$$

$$\begin{bmatrix} O \\ Y \end{bmatrix} = [0 \; 0 \quad 0 \; 0 \quad y_1{}' e_1{}' y_2{}' e_2{}' \ldots]'.$$

Here the $y_i{}'$ represent guesses; and the $e_i{}' = y_i - y_i{}'$ represent not only a set of error terms in supervector g but also a set of residual ordinates to be included in the supervector h for the assurance of "closure"! The errors, $e_i{}'$, are actually <u>known</u> in this case. The observed y_i do <u>not</u> enter as such into h. The increments a" and b" are readily computable, together with the a' and b' corresponding to the guessing round, in a single pass.

A more striking pattern may be obtained by a rearrangement of terms in the design supermatrix and in the two supervectors, as in Figure 1.

$$
\left[\begin{array}{cc}
\begin{array}{c|c} 0 & \begin{matrix}1 & \cdots \\ x_1 & \cdots\end{matrix} \\ \hline \begin{matrix}1 & x_1 \\ \vdots & \vdots \\ \vdots & \vdots\end{matrix} & \begin{matrix}1 & & \\ & \ddots & \\ & & \ddots\end{matrix}\end{array} & 0 \\ \hline
0 & \begin{array}{c|c} 0 & \begin{matrix}1 & \cdots \\ x_1 & \cdots\end{matrix} \\ \hline \begin{matrix}1 & x_1 \\ \vdots & \vdots \\ \vdots & \vdots\end{matrix} & \begin{matrix}1 & & \\ & \ddots & \\ & & \ddots\end{matrix}\end{array}
\end{array}\right]
\cdot
\left[\begin{array}{c} a' \\ b' \\ e_1' \\ \vdots \\ \vdots \\ a'' \\ b'' \\ e_1'' \\ \vdots \\ \vdots \end{array}\right]
=
\left[\begin{array}{c} 0 \\ 0 \\ y' \\ \vdots \\ \vdots \\ 0 \\ 0 \\ e_1' \\ \vdots \\ \vdots \end{array}\right]
$$

Fig. 1

Notice that the leftmost expression is a "supermatrix of supermatrices," replicating the F required for classical unweighted least squares. This form is of interest for the next example and for our brief remarks on multi-equation models.

My second example shows the instant supermatrix setup for "stepwise" regression, a concept due to Goldberger and others [7], [8]. A linear relationship is estimated progressively for one set of error-affected y's from two or more sets of x's. The procedure is usually presented for only two sets of x's but additional sets can be accommodated easily within the supermatrix framework.

For two steps, the supermatrix system appears as in Figure 2.

$$
\left[\begin{array}{cc}
\begin{array}{c|c} 0 & \begin{matrix}1 & \cdots \\ x_1 & \cdots\end{matrix} \\ \hline \begin{matrix}1 & x_1 \\ \vdots & \vdots \\ \vdots & \vdots\end{matrix} & \begin{matrix}1 & & \\ & \ddots & \\ & & \ddots\end{matrix}\end{array} & \\ \hline
 & \begin{array}{c|c} 0 & \begin{matrix}1 & \cdots \\ x_1^* & \cdots\end{matrix} \\ \hline \begin{matrix}1 & x_1^* \\ \vdots & \vdots \\ \vdots & \vdots\end{matrix} & \begin{matrix}1 & & \\ & \ddots & \\ & & \ddots\end{matrix}\end{array}
\end{array}\right]
\cdot
\left[\begin{array}{c} a \\ b \\ e_1 \\ \vdots \\ \vdots \\ a^* \\ b^* \\ e_1^* \\ \vdots \\ \vdots \end{array}\right]
=
\left[\begin{array}{c} 0 \\ 0 \\ y_1 \\ \vdots \\ \vdots \\ 0 \\ 0 \\ e_1 \\ \vdots \\ \vdots \end{array}\right]
$$

Fig. 2

Here, the y_i are observed ordinates and the e_i are residuals -- the differences between the observed ordinates and the computed estimates derived from the first regression step. The second step involves, in effect, the regression of the first residual ordinates on the x_i^*, a second set of observations. Since the e_i are not known in advance in this case (i.e., they result from computation, not from guessing), the solution of the supermatrix system cannot be accomplished in one pass.

Multi-equation Models

More ambitious applications of instant supermatrix design are suggested by a glance at texts exhibiting: the generalized least-squares method of Aitken, the three-stage approach to least-squares estimation of Zellner and Theil, and the discrete-time linear engineering formulations that accommodate "transitions" with "noise" from an initial state to subsequent ones [9], [7], [8]. In these instances, which involve the simultaneous estimation of all the parameters in a multi-equation model, the supermatrix setup is introducible in the last phases. The column of disturbances is absorbed into the supermatrix equation Fg = h. The elements of F and h would commonly have undergone substantial processing rather than represent raw data.

The two examples discussed in the preceding section also suggest analogues in which multiple equations are treated simultaneously. In these analogues, the errors arising in the first round or step are not introducible as data in successive ones.

Another possible application of the supermatrix approach is to instances in which a full linear system is to be estimated from observations made on fragmentary linear "shards." All the y_i may, for example, represent intelligence quotients or performance scores of a given kind. Associated with a subset of the y_i are ratings in one or more "predictive" tests; associated with other subsets of y (they may differ in number of elements) are ratings in still other tests or in various combinations of tests. A supermatrix system can be designed immediately for the derivation of the single least-squares equation accommodating all of the data simultaneously. This equation may be useful in its own right or for comparison with the regression estimates corresponding to y_i subsets.

Finally, it appears feasible to estimate by supermatrix methods the unknowns of a multi-equation model that consists of (a) accounting identities presumed to be exact and (b) inexact regression relationships. The latter equations may, of course, include highly processed "observations." They may express each of several endogenous variables in terms of exogenous variables, other endogenous variables, or both [10].

REFERENCES

[1] Siegel, I.H., "The Supermatrix Approach to Least-Squares Adjustment," The American Statistician, December 1968, 22-23.

[2] ____________, "Instant Matrix Design: Straight Line With Errors in Both Variables," Proceedings of the Social Statistics Section, American Statistical Association, 1968, 247.

[3] ____________, "Difference Between Paasche and Laspeyres Index-Number Formulas," Journal of the American Statistical Association, September 1941, 343-350.

[4] ____________, "Further Notes on Index-Number Differences," ibid., December 1941, 519-524.

[5] ____________, "Simplified Least-Squares Approach for an Age of Computers," Proceedings of the Social Statistics Section, American Statistical Association, 1966, 398-400.

[6] ____________, "From Identities to Normal Equations: An Easy Approach to Least-Squares," Proceedings of the Social Statistics Section, American Statistical Association, 1967, 354-356.

[7] Goldberger, A.S., Econometric Theory. New York: Wiley, 1964.

[8] Deutsch, R., Estimation Theory. Englewood Cliffs: Prentice-Hall, 1965.

[9] Christ, C.F., Econometric Models and Methods. New York: Wiley, 1966.

[10] Fox, K.A., Sengupta, J.K. and Thorbecke, E., The Theory of Quantitative Economic Policy. Chicago: Rand McNally, 1966.

POSTSCRIPT

Since it is my nature to abhor a vacuum, I shall use some of the abundant space that remains on this page for a few additional remarks on sequential estimation.

First, the example I used for stepwise regression was deliberately patterned on the example preceding it and hence may not at first resemble the illustrations usually given. The supermatrix approach is adaptable, of course, to instances in which the second set of observed independent variables does not, say, include a column of 1's.

I wish to point out also that the classical cases of least-squares adjustment may readily be restated to involve sequential procedures. Indeed, the example that starts with a guess suggests one of the many routes that might be taken toward enlargement of the standard design supermatrices without alteration of the final results.

Another route is suggested by a theorem published by Jacobi in 1841 (it is shown in Whittaker and Robinson, Calculus of Observations). This theorem, which may be translated into supermatrix form at once, says that the unweighted least-squares straight line is equivalent to a certain weighted average of the ${}^{n}C_2$ lines derivable from n given values of x and y.

Still another way is to introduce pairs of mathematically redundant unit submatrices into the design supermatrix for the unweighted case. These diagonal matrices are like "idlers" or "gears." The first pair would link the residuals to a set of Lagrangian coefficients; the second pair would link these coefficients to another set of multipliers; and so on.

The weighted case is more interesting. Here, too, pairs of extra diagonal matrices may be inserted into the design supermatrix, one diagonal showing the weights and the other showing 1's. If two or more tiers are introduced, the weights included in each may be raised to a fractional power; but the introduction of tiers must cease when the sum of the fractional powers reaches unity.

I conclude with the comment that an unweighted least-squares system may also be translated into a larger weighted one in which Q does not represent the transpose of P. In this setup, a tier containing a compensatory matrix of pseudo-weights and a diagonal unit matrix is included in the design supermatrix. The matrix of pseudo-weights, however, is not diagonal and not unique, and it requires some computation.

ASSESSING THE ACCURACY OF ESTIMATES: A PERSONALIST APPROACH

Rex V. Brown, University of Michigan

Introduction

The orientation of this talk is strictly practical, in the sense that the ultimate beneficiary is intended to be the man of affairs, like a businessman, who may use the fruits of the statistician's labors, rather than the statistician himself. If the approach to be presented does not contribute to the technology of business administration and other applied arts, it has failed in its primary purpose. If it happens also to have some academic interest, so much the better.

At a theoretical level, however, there is nothing very original in the approach. It does, however, appear to represent a novel attack, at an operational level, on a scandalously mishandled problem in decision-oriented quantitative research of all kinds. This problem is a trite and universal one for anyone who has to make decisions in the face of uncertainty; and it is in two parts:

1. How to assess uncertainty about relevant target variables (such as a market share), which I will call the problem of Target Assessment.
2. How to evaluate ways of reducing this uncertainty, which I will call the problem of Research Design.

Now, I am well aware, of course, that the literature abounds with procedures that appear to address these problems. In the face of sample findings, for example, such devices as Confidence Intervals, Maximum Likelihood Estimates, and Tests of Significance, certainly seem to be saying something about Target Assessment. However, the trouble with classical inference tools such as these is that their output is not in a form that is of direct interest to a decision maker. He wants to answer the very personal question, "Where does my target variable probably lie?", whereas a Confidence Interval, for example, is telling him how surprising the observed research results would be if some variable, (not necessarily his target variable) had some hypothetical values.

When he considers a possible Research Design he wants to answer the question, "What research can I do which will make me least uncertain?", rather than, "What research will produce the smallest sampling variance, from among those research designs for which a sampling variance can be objectively calculated?" The weaknesses of classical inference for decision-oriented purposes are well known and do not need to be covered in any detail here.[1]

When assessing his target variable, the decision maker surely wants to come up with a personal probability assissment, possibly in form shown in Figure 1. However, he may not care to represent his assessment as a complete probability or credence distribution (as I prefer to call a personal probability distribution when it refers to an assessment conditional only to what the assessor actually knows).[2] When choosing between research designs, he will also surely want to look ahead to the kind of probabilistic assessment he will make after the research. Presumably, he will opt for the design which, in some sense, promises to produce a personal probability distribution with as little "spread" as possible.

Personalist Inference

What is commonly known as Bayesian inference is, of course, designed to handle exactly this kind of problem and substantial development in the area has been achieved, notably by Savage, Raiffa, and Schlaifer. However, not very much of this development has yet trickled down into real world usage.

I have recently completed a brief survey of what use business is making of Bayesian decision theory. While I found plentiful and even dramatic usage of decision trees and other devices for analyzing decisions given specified assessments of uncertainty about the relevant target variables, I found almost no use of the conventional Bayesian devices for determining those uncertainties. I am referring, of course, to prior-posterior analysis and its derivative preposterior analysis. At least, I was able to find very few instances where executives acted on the implications of such Bayesian analysis.

Now, part of this lack of implementation is no doubt due to the quite natural time lag between a new technology being developed at a theoretical level, and its becoming operational. But part, I think, is because the technique itself, as currently developed, is not always appropriate for use by non-technical decision makers or, as I will call them, executives.

Both prior-posterior and preposterior analysis depend upon the use of Bayes' Theorem and require two critical inputs which the executive has, if not necessarily to supply, at least to agree with. In the first place, a prior distribution must be assessed on the target variable and it may be exceedingly difficult to make an uncontaminated prior assessment after research evidence has already been obtained. Commonly, this is the point in time at which a prior posterior analysis will be initiated. Secondly, a likelihood function must be assessed, specific to actual or potential research findings, and it may be very difficult for a non-technical executive to grasp what it means, let alone contribute to its formulation. Except in those rare cases where the data generating process, conditional on the true value of the target variable, is uncontroveraially known (e.g., random sampling with perfect measurement), informed judgment needs to go into the construction of a Likelihood Function. The most appropriate judgment will often be in a head unable to express it in a Likelihood Function form.

In addition to these difficulties of eliciting needed inputs, I have found that very few executives feel that they understand even the general idea of prior-posterior or preposterior inference. For this reason they are understandably hesitant to trust decisions that may involve millions of dollars to an analysis based on an arcane logic.

Is there any way of avoiding these drawbacks? I think there is, and I would now like to propose an alternative which, while it is Bayesian, in the sense that it accepts personal inputs and its output is interpreted personally, it does not depend on Bayes' Theorem. (I think it would be a very good convention if we could agree to use the word "Bayesian" only for those types of personalist analysis which depend on Bayes' Theorem.)

Credence Decomposition

The alternative I am about to propose depends, not on Bayes' Theorem, but on the equally well known logic of the distribution of functions of random variables. I call it Credence Decomposition and, in the form I shall now present, it can be used both for problems of Target Assessment and Research Design.

The essential steps are very simple and are as follows:

1. The target variable is decomposed, by which I mean that it is expressed as a function of two or more components. A very simple example would be to express future sales of a product as sales per outlet times number of outlets. A slightly more elaborate decomposition (and decompositions can get very elaborate) would be to express future sales as the sum of multiplicative expressions of the above form for each of a number of market sectors.
2. Each component thus defined is assessed probabilistically on whatever evidence is available to the assessor. This could include field work, judgment or published statistics and the supporting reasoning could be any combination of intuition and statistical theory (including, possibly, prior-posterior analysis).
3. A personal probability distribution, e.g., in the form of Figure 1, is derived routinely by any of a number of standard statistical procedures. Computer programs, formulas and other supporting devices have been developed to make this processing as painless as possible.

At this level of generality, Credence Decomposition is a rather trivial (if unexploited) tool. However, there is a variant of Credence Decomposition which is less obvious and which seems to lend itself rather conveniently to problems of target assessment and Research Design. I call this Error Decomposition.

Error Decomposition

In Error Decomposition, what is decomposed is not the target variable of ultimate interest to the executive but the estimating error resulting from a specific piece of quantitative research, e.g., a sample survey.

There are at least two ways of formally defining estimating error. It can be defined as the difference between the target variable and some more or less arbitrary estimate calculated from the research findings. Alternatively, it can be the ratio of the target variable to such an estimate. Either formulation has advantages in different circumstances, and for illustrative purposes, I will discuss only the error ratio form.

Let me take as a concrete example a parking survey situation that has been written up in some technical detail elsewhere.[3]

The case setting is as follows: A British town, which I will call Camford, had a mail survey done to assess the probable demand for parking space if meters were introduced. A list of ten thousand locally registered motorists was obtained and, of these, one thousand were randomly selected and sent a mail questionnaire. Nine hundred returned the questionnaires and of these ten percent or ninety indicated that, if meters were introduced, they would be parked in the downtown area at a given peak hour.

The city engineer's target assement problem is, what should he conclude personally about the actual demand for parking space if meters were introduced, expressed in a form like Figure 1? He also has a research design problem. If he conducts a new survey in another town, should he use the same budget on another mail survey or on a smaller personal survey.

On the target assessment problem, the first thing the city engineer might do, using Credence Decomposition, would be to decompose the total spaces needed (if meters were introduced) as the product of:

> The fraction of local motorists needing space (t)
> TIMES
> Number of local motorists (n)
> TIMES
> Some adjustment factor to allow for spaces needed by out-of-town parkers (f)

If probabilistic assessments can be made for each of these, a credence distribution on the target variable can be derived routinely. The number of local motorists (n) is known to be ten thousand, so no probabilistic assessment is needed of that component. The out-of-town adjustment component (f) can be assessed by direct intuition informally. This leaves the "local fraction," (t), which is the variable which the mail survey addresses. This is the assessment he might use error decomposition for.

Figure 2 shows the essential steps the assessor might go through, viz.: to express total error ratio as a function of component ratios which reflect distinguishable (and assessable) sources of error. The nested rings at the top of the figure and the vertical lines indicate the various ways in which sources of error can creep in between the true value of the target t, "local fraction," and the estimate a' (known to be ten percent).

t/a' is then the total error ratio and the component error ratios are defined in the line in Figure 2 marked "Decomposition." It can be seen that three sources of error are distinguishable: random error, nonresponse error, and reporting error. It can easily be verified that each of these will be one if there is no error of the type involved. The set of boxes on the right hand side of the bottom line of Figure 2 summarize, in the form of ninety-five percent credible intervals, probabilistic assessments that were made for each of the three component error ratios.

The detail of these assessments and the logic behind them are described in the reference cited in Footnote 3. Suffice it to say here that the random error was based on prior-posterior analysis using a flat prior assumption, and the other two errors were assessed intuitively. The resulting distribution of t/a' was approximated by means of a formula which exploits the fact that the relative variance of a product is approximately the sum of the relative variances of the components. (Also discussed more fully in the above reference.) It could also be computed more exactly by a computer program called DECOMP.[4]

The credence distribution on t follows directly from the distribution on t/a' and, as a summary, its credible interval appears in the left hand box in Figure 2.

The city engineer might thus conclude, if he accepts the input assessments, that he can be ninety-five percent sure that the local fraction, t, lies between six percent and twenty-one percent. Conjoined with the knowledge that there are ten thousand local motorists and an assessment of the "out-of-town adjustment" with a credible interval of 1 to 1.2, a credence curve on the real target variable, total spaces needed, was derived, which actually corresponds to Figure 1. His target assessment would therefore be that between three hundred and twenty-two hundred parking spaces will be needed if meters are introduced, with ninety-five percent personal probability.

As for the research design problem, the city engineer would go through virtually the same procedure for each of the alternative research designs considered. If the cost is the same, he might reasonably choose which ever strategy produces the smallest span for the credible interval on the total error ratio. (Alternative research design criteria can be selected, such as prior expectation of posterior variance, but they seem to produce almost identical rankings.)

Conclusion

There has only been time this afternoon to rather briefly sketch the class of algorithms which I call Credence Decomposition. A fuller technical discussion appears in the article referenced in Footnote 3. A general handbook on the technique and its applications is in process of publication.[5]

The general Credence Decomposition technique and the Error Decomposition variant of it seems to work reasonably well in the Marketing Research area in which I am most familiar. At the present time, it certainly seems to suffer from very little competition as far as practical tools for research appraisal are concerned.

Other researchers, notably Professor Charles Mayer of York University, are working on the critical problem of how to make reasonable, empirically based component assessments which are required by this technique or others with the same objectives.[6] I would certainly appreciate hearing from anyone else who may be working in this area.

FOOTNOTES

1 See J. W. Pratt, H. Raiffa, and R. Schlaifer, "Introduction to Statistical Decision Theory," McGraw Hill, 1965, Chapter 20.

2 Thus a sampling distribution conditional on some hypothetical value of a population parameter would not be a credence distribution though it might well be a personal probability distribution.

3 See Rex V. Brown, "The Evaluation of Total Survey Error," The Statistician, Volume 17, No. 4, 1967, (copies of this paper are available from the author).

4 Further information about this program can be obtained from the author.

5 Rex V. Brown, "Research and the Credibility of Estimates," Harvard University, Graduate School of Business Administration, Division of Research, Boston, 1969.

6 Charles S. Mayer, "Assessing the Accuracy of Marketing Research," to be published.

XII

CONTRIBUTED PAPERS II

FAMILY AND PERSONAL INCOME: MEASUREMENT PROBLEMS AND FINDINGS

Chairman, EVA MUELLER, University of Michigan

INCOME NONRESPONSES IN THE CURRENT POPULATION SURVEY*

Mitsuo Ono and Herman P. Miller
U.S. Bureau of the Census

Introduction

Since 1947, the Bureau of the Census has collected annual income information in the March supplement to the Current Population Survey (CPS). As in all household sample surveys which rely on voluntary cooperation of sample units to participate in the survey, one of the more persistent problems encountered in the March CPS has been the need to minimize the bias resulting from noninterviews and nonresponses to income questions.1/ Noninterviews represent failures either to contact designated sample units or to secure their cooperation when contacted. Nonresponses refer to partial or complete failures to secure information on questionnaires from respondents. Income information is obtained by interviewers asking eight different questions for each person 14 years old and over living in the sample unit and failure to obtain information on any one of these eight questions for a person makes him a nonrespondent.2/ It is also noted that a nonresponse does not necessarily mean that the entry should have been either a positive or negative amount. Many of the nonresponses can be "none's." The bias from missing income information in these surveys results from the fact that nonrespondents usually have different economic and social characteristics than respondents. Nonrespondents tend to have higher incomes, are classified in higher skilled occupations, and work at full-time jobs. Although many devices are used in the CPS to reduce income nonresponses, missing income information in the CPS continues to be an important problem. As table A-1 shows, the percent of families in which there were one or more income items not reported (NA) for any family member 14 years old and over has increased from 11 percent for survey year 1960 to about 19 percent for survey year 1969. When the noninterview rate of about 5 percent is added to this family income nonresponse rate, it is estimated that currently complete income information is not obtained for about one-fourth of all families covered in the CPS.

This paper covers various aspects of the income nonresponse problem in the CPS. It outlines procedures used to impute missing income information in the CPS, shows characteristics of persons associated with relatively high income nonresponse rates, examines the overall impact of allocation procedures on income data, and outlines changes in the income collection procedures in the March 1969 CPS.

Treatment of Income Nonresponses in the CPS

Income nonresponses can be treated in three ways. The first method involves publishing in detail only reported income information and showing the "not reported" information as a single line. The second method is to publish data in which fully reported data were inflated to control totals. This procedure assumes that characteristics of respondents and nonrespondents are alike. The third method involves imputing income values for missing information based on data obtained from persons with similar social and economic characteristics who reported completely their income information.

Before the 1962 CPS, the publication of income data showed only units which reported fully on income. The one exception to this general rule was that if a person had a combined income of $10,000 or more from all sources reported, missing income entries were treated as zeroes. Beginning with the 1962 CPS, missing income data were "allocated" in such a way that when a person did not answer one or more of the income items, all of his income data were replaced by information obtained from a person with similar economic and demographic characteristics, e.g., age, sex, color, occupation, work experience, etc. The income amounts for the last person with certain social and economic characteristics who provided complete information were stored in specific locations in the computer. When a nonrespondent appeared, the income information for the last person with similar characteristics was substituted. Beginning with the March 1966 CPS, the "allocation" procedures were further improved. First, instead of assigning completely new income information for nonrespondents who had one or more missing income values, only that type of income which was not reported was allocated. However, the "imputation" procedure was still based on reported income values of persons with similar social and economic characteristics. Second, the "$10,000 and over" rule was eliminated so that missing income information for all nonrespondents was imputed, regardless of income levels. Third, before the March 1966 CPS, there was no regular procedure for allocating work experience to persons who were in the same rotation groups in both the February and March supplements but who could not be matched. The tables published in the P-60 reports relating to work experience, occupations, and industry of longest job were limited only to those persons who could be matched in the February and March Supplement CPS surveys. However, a special analysis showed that matched and nonmatched persons tend to have different characteristics. Thus, there were more nonmatched cases in nonfarm areas and among women or persons classified other than white. Consequently, an allocation procedure was initiated to assign nonmatches with work experience information from fully reported respondents with similar social and economic characteristics. Estimated values were used only when there was a nonmatch case within a given classification, but no respondent with similar social and economic characteristics. Fourth, the allocation matrix was expanded. An example of the revised procedure is shown in the following table.

*Statistical assistance by Messrs. Emmett Spiers and Joseph Knott, Population Division, and Dr. Richard Hornseth, Systems Division, is gratefully acknowledged.

Persons' characteristics	Wage or salary income	Nonfarm self-employment income	Farm self-employment income	Income other than earnings
Completely Reported Person				
White male, 25 to 34 years, worked 52 weeks, self-employed professional..........	None	$8,000	None	$100
Not Reported Person				
White male, 25 to 34 years, worked 52 weeks, self-employed professional..........	None	NA	None	None
Allocated Person				
White male, 25 to 34 years, worked 52 weeks, self-employed professional..........	None	$8,000	None	None

Thus, assume that a self-employed white male professional, age 25 to 34 years old and who worked 52 weeks reported that he had no income except that amount received from his professional work, but he did not report this information. Only this amount (only item missing) would then be allocated from a respondent with similar characteristics with the other items still remaining the same.

As in past procedure, the "good" income values assigned to a nonrespondent are those stored for the last person with similar social and economic characteristics and who had been selected systematically in the order in which individual records were processed. However, when there was a nonresponse case and there was no respondent with similar social and economic characteristics, estimated initial values were used.

Some Characteristics of Nonrespondents

The income nonresponse rate is computed by dividing the number of nonrespondents (defined as units with one or more missing income entries) by the total number of units after allocation.

Units in higher income intervals tend to have higher income nonresponse rates. Thus, for families in the 1969 CPS, the income nonresponse rates were 24 percent and 35 percent in the $15,000 to $24,999 and $25,000 and over intervals, respectively. The income nonresponse rate was also relatively large in the $1 to $999 or loss income interval but this class contains only about 1.5 percent of all families. (See tables B-1 and B-2.)

Data for the survey year 1967 covering 1966 income (the most current information available) show that overall, men had relatively higher income nonresponse rates than women. However, female unrelated individuals had higher income nonresponse rates than male unrelated individuals. White persons had higher income nonresponse rates than Negro persons and others. Persons who worked at full-time jobs tend to have higher income nonresponse rates than persons who worked at part-time jobs. Among men, persons classified as either professional, technical and kindred workers (21 percent) or managers, officials, proprietors excluding farm managers (28 percent) had relatively higher income nonresponse rates than others. The highest income nonresponse rates (42 percent and 34 percent) were associated with self-employed men in these two occupational categories. Among women, the highest income nonresponse rates were also related to the self-employed. (See table C.)

Table D shows income nonresponse rates by types of income. For both men and women, the highest income nonresponse rates were related to "other income" types, e.g., Social Security, property income, public assistance, etc.

Table E attempts to show the pattern of income nonresponses for household heads. The unweighted nonresponse data (obtained from a special sample tally of March 1968 CPS questionnaires) show that income nonresponses tend to fall predominantly in two categories: Across-the-board blanks (30 percent) and one missing source only (45 percent). Data also show that missing income information on property income (interest, dividends, and net rental income) was associated with approximately 333 household heads or about 61 percent of the 548 household heads who had at least one missing income item.

Impact of Allocation Procedures

Tables B-1 and B-2 show income distributions of respondent units (before allocation), nonrespondent units, and the combination of respondent and nonrespondent units (after allocation) for the March 1969 and 1968 CPS. The median income of nonrespondent families is about $840 higher than that of respondents and the net impact of the allocation procedures is to raise the median income for the entire distribution by about $140 (from $8,547 before allocation to $8,689 after allocation).

These tables do not provide any information on the number of families that had complete blanks on income information or on the amount of income that was allocated. Assume, for example, that a family reported wage and salary income of $12,000, but did not report interest and dividends and that $100 of interest and dividends was imputed. This family would then be tabulated as a nonresponse at the $10,000 to $14,999 income class despite the fact that nearly all of the income for that family was reported. Consequently, the sample nonresponse rate may not accurately reflect the impact of nonresponse on aggregate family income. A special analysis was made for one-fourth of the March 1968 CPS sample (two out of eight sample rotation groups) in order to examine this problem. Data in table F show that about 18 percent of nonrespondent families accounted for only 10 percent of aggregate income. Overall, the impact of nonresponse on the CPS income data is considerably less than the nonresponse rate suggests. The ratio of allocated aggregate income to total aggregate income can be called the allocated income share rate. The income nonrespondent rate and the allocated income share rate varied by different family income brackets. Thus, higher nonresponse rates were positively associated with larger family incomes (excluding families with no income which accounted for about a third of a percent of all families). This pattern was also found for the allocated income share rate. This rate increased from 8 percent in the under $3,000 family income bracket to 13 percent in the $15,000 to $24,999 family income bracket. However, although the nonresponse rate within the $25,000 and over bracket was relatively high (33 percent), the allocated income share rate was only about 9 percent. For families in the $15,000 to $24,999 family income bracket, the nonresponse rate was 26 percent while the allocated income share rate was about 13 percent. These data suggest that large numbers of families at the higher income levels report their main source of income but omit relatively minor amounts received from interest, dividends, and similar sources. The data also imply that nonrespondent families within the higher income brackets can be considered mostly "partial" nonrespondents and not "complete" nonrespondents (all income items blank).

Changes in Methods of Collection Income Data in the 1969 Current Population Survey

For the March and April 1969 CPS Supplements, the Bureau of the Census introduced modifications in the collection of income data. A major change in the income collection method previously used in these CPS supplements was the extension of the interview period for the collection of income data in the six rotation groups in which income questions are asked in March. Interviewing for labor force and income information is done during the calendar week including the 19th of the month. The rigid time schedule required for the publication of the labor force data requires that all schedules be returned to Washington early the following week. This early closeout presumably accounts for some part of the relatively high nonresponse rates because many household respondents, who are often housewives, would not know about the income amount received by other family members, although they will know about the labor force status of other family members, e.g., who had worked or not during the preceding week, the kind of work they did, and the number of hours worked. A new followup form was designed to accommodate those household respondents who need more time to obtain the required income information. It is estimated that some member of approximately 4,400 households or 9 percent of the total applicable households made use of these followup forms.

In addition, several improvements were made in the March 1969 CPS. These include: (1) Modification of the design and content of the questionnaire to allow for more detailed questioning of the income items; thus, boxes for gross receipts, business expense, and net income have been added to self-employment earnings questions to help the interviewer and respondent determine net income; (2) field office editing procedures were extended to a 100-percent income edit of the CPS schedules; (3) training instructions were strengthened by including more detailed explanations and more examples; and (4) the interview group training session was shifted to March (from February).

In addition to the above, a research program to determine better methods of collecting income data was instituted. This program was designed to answer the following questions:

1. Does the collection of income data in April, rather than in March, result in more complete information? The income data collected after the taxfiling date in April are to be compared against data collected in other periods with respect to nonresponse rates and income levels. One rotation group consisting of about 6,000 households was used for this purpose.

2. Does the collection of income and work experience data within a single interview result in more accurate information? At present, information on work experience (e.g., whether worked or not last year, occupation and industry last year, etc.) is collected in February and April and the information on income received last year is collected in March and the records are merged in the computer to generate data cross-classifying income received with work experience information for the preceding year. As might be expected there are many inconsistencies in reporting, e.g., some people reporting work experience in February but reporting no income in March, etc. These inconsistencies are currently adjusted by a fairly elaborate computer allocation procedure. If large improvements are obtained by collecting both items of information in a single interview, it may be possible that these computer matching inconsistencies will not only be eliminated but the collection of these complementary data would improve the quality of both data. This procedure, however, results in a much lengthier interview and the possibility of higher nonresponse rates in this or subsequent surveys. Thus, it is planned to analyze both benefits and cost of implementing these procedures. One rotation group was used to explore this specific problem.

3. Does the advance notification to the respondent of the types of income questions to be asked result in more accurate information? For February, all of the households in one rotation group were sent a letter requesting cooperation

and including income questions that were to be asked in March. It was hypothesized that the advance letter would elicit more cooperation from the respondent in reporting income data.

Preliminary findings of 1969 CPS income nonresponse data show that despite the efforts to improve response, the income nonresponse rate increased by about 2 percentage points overall over the preceding year. Refusals to provide any income information appear to account for a large portion of the nonresponses. Also, a preliminary evaluation of the 1969 data suggests that aggregate income in the CPS relative to independently derived benchmark data increased about 2 percentage points (from 85 to 87) over the preceding year. Most of the improvement appears to be in property income. These findings are still tentative inasmuch as the preliminary data have only recently become available. A more intensive examination of the results of the experimental program is now underway.

FOOTNOTES

1/ This document covers mainly topics on the income nonresponse problem. It is noted, however, that in the March 1969 CPS, no information was recorded for approximately 5 percent of the 50,000 occupied households because no interviews could be obtained during the week in which the enumeration was conducted.

2/ In the March 1969 CPS, the eight income questions covered: Wage and salary; nonfarm self-employment income; farm self-employment income; Social Security or Railroad Retirement; dividends, interest, net rental income, or income from estates or trusts; public assistance or welfare payments; unemployment compensation, government employees pensions or veterans' payments; and private pensions, annuities, etc.

Table A-1.--INCOME NONRESPONSE RATES IN THE CPS (1948-1969) FOR FAMILIES AND UNRELATED INDIVIDUALS, BY COLOR OF HEAD, FOR THE UNITED STATES

Survey year	Families			Unrelated individuals		
	Total	White	Negro and other	Total	White	Negro and other
March 1969 CPS (1107)...	19.0	19.2	17.4	14.5	14.8	11.9
March 1968 CPS (1107)...	17.2	17.3	16.2	12.6	13.1	9.9
March 1967 CPS (1107)...	21.9	22.3	18.0	15.5	16.3	10.1
(1105)...	19.0	19.5	15.1	17.2	17.6	14.4
March 1966 CPS (1105)...	14.8	15.0	12.8	15.4	15.5	14.7
March 1965 CPS (1105)...	14.0	14.2	12.7	13.9	14.0	13.5
March 1964 CPS (1105)...	NA	NA	NA	NA	NA	NA
March 1963 CPS (1105)...	NA	NA	NA	NA	NA	NA
March 1962 CPS (1105)...	NA	NA	NA	NA	NA	NA
March 1961 CPS*.........	9.0	9.1	7.3	11.8	11.6	13.0
March 1960 CPS*.........	10.5	10.4	10.8	12.5	12.3	13.8
March 1959 CPS*.........	10.9	11.0	10.4	13.0	12.2	17.7
March 1958 CPS*.........	11.2	11.4	8.6	12.0	12.2	10.7
March 1948 CPS*.........	7.5	7.9	5.3	7.3	7.6	5.2

* Prior to March 1962 CPS income nonresponses were not allocated.
NA - Not available.

Table A-2.--INCOME NONRESPONSE RATES IN THE CPS (1948-1969) FOR PERSONS 14 YEARS OLD AND OVER, BY COLOR AND SEX, FOR THE UNITED STATES

Survey year	Male			Female		
	Total	White	Negro and other	Total	White	Negro and other
March 1969 CPS (1107)...	15.0	15.2	13.3	12.2	12.4	10.3
March 1968 CPS (1107)...	12.7	12.9	11.4	9.7	9.7	9.4
March 1967 CPS (1107)...	16.2	16.6	12.6	11.9	12.2	9.6
(1105)...	15.2	15.6	11.3	9.3	9.5	7.3
March 1966 CPS (1105)...	NA	NA	NA	NA	NA	NA
March 1965 CPS (1105)...	10.9	NA	NA	6.9	NA	NA
March 1964 CPS (1105)...	10.5	NA	NA	6.6	NA	NA
March 1963 CPS (1105)...	9.1	NA	NA	5.6	NA	NA
March 1962 CPS (1105)...	8.9	NA	NA	5.3	NA	NA
March 1961 CPS*.........	7.7	7.8	6.9	4.4	4.6	2.9
March 1960 CPS*.........	9.0	8.9	9.6	5.0	5.0	5.3
March 1959 CPS*.........	9.3	9.2	9.4	5.0	5.0	5.0
March 1958 CPS*.........	9.2	9.5	7.2	4.9	5.0	4.1
March 1948 CPS*.........	6.7	NA	NA	3.8	NA	NA

* Prior to March 1962 CPS income nonresponses were not allocated.
NA - Not available.

Table B-1.--FAMILIES AND UNRELATED INDIVIDUALS BY TOTAL MONEY INCOME IN 1968 AND 1967, BEFORE AND AFTER ALLOCATION OF INCOME NONRESPONDENTS, FOR THE UNITED STATES

Total money income	Families								Unrelated Individuals							
	March 1969 CPS				March 1968 CPS				March 1969 CPS				March 1968 CPS			
	After allo-cation	Before allo-cation	Non-re-spond-ents	NA rate	After allo-cation	Before allo-cation	Non-re-spond-ents	NA rate	After allo-cation	Before allo-cation	Non-re-spond-ents	NA rate	After allo-cation	Before allo-cation	Non-re-spond-ents	NA rate
Number...thousands...	50,510	40,921	9,589	19.0	49,834	41,279	8,555	17.2	13,803	11,808	1,995	14.5	13,114	11,458	1,656	12.6
Percent.........	100.0	100.0	100.0	(x)	100.0	100.0	100.0	(x)	100.0	100.0	100.0	(x)	100.0	100.0	100.0	(x)
None.................	0.3	0.3	0.2	15.9	0.4	0.3	0.6	29.2	2.5	2.5	2.3	13.3	3.1	2.9	5.0	20.2
$1 to $999 or loss...	1.5	1.4	1.9	23.4	1.7	1.7	1.8	18.0	12.4	12.2	13.3	15.6	16.4	16.1	18.7	14.4
$1,000 to $1,999.....	3.4	3.6	2.8	15.7	4.4	4.4	4.4	17.3	24.5	24.4	24.7	14.5	24.8	24.7	25.5	13.0
$2,000 to $2,999.....	5.1	5.3	4.3	16.0	6.0	6.2	4.9	14.0	12.8	12.6	13.6	15.3	12.7	12.9	11.4	11.3
$3,000 to $3,999.....	6.1	6.2	5.5	17.1	6.3	6.5	5.6	15.3	10.9	11.0	10.6	14.0	9.7	9.7	10.1	13.1
$4,000 to $4,999.....	6.0	6.1	5.7	17.8	6.5	6.7	5.5	14.6	8.0	8.0	8.4	15.2	7.6	7.6	7.8	13.0
$5,000 to $5,999.....	6.9	7.0	6.2	17.2	7.8	8.0	6.8	15.0	6.7	6.9	5.3	11.5	7.1	7.2	6.0	10.6
$6,000 to $6,999.....	7.6	7.7	6.8	17.1	8.3	8.5	7.2	14.9	5.3	5.6	3.5	9.6	5.5	5.7	4.3	9.9
$7,000 to $9,999.....	23.4	23.9	20.9	17.0	24.2	24.8	21.4	15.2	10.2	10.2	10.0	14.1	8.0	8.3	6.0	9.6
$10,000 to $14,999...	25.0	24.9	25.4	19.3	22.4	22.2	23.4	17.9	5.0	4.9	5.7	16.5	3.4	3.5	3.1	11.6
$15,000 to $24,999...	12.1	11.4	15.4	24.1	9.6	8.7	13.9	25.0	1.3	1.2	2.1	23.2	1.2	1.2	1.6	16.2
$25,000 and over.....	2.6	2.1	4.9	35.2	2.4	2.0	4.4	30.9	0.4	0.4	0.6	18.7	0.5	0.5	0.5	13.9
Median income........	$8,689	$8,547	$9,383	(x)	$8,067	$7,928	$8,842	(x)	$2,838	$2,860	$2,720	(x)	$2,450	$2,499	$2,070	(x)

x - Not applicable

Table B-2.--PERSONS 14 YEARS OLD AND OVER BY TOTAL MONEY INCOME IN 1968 AND 1967, BEFORE AND AFTER ALLOCATION OF INCOME NONRESPONDENTS, BY SEX, FOR THE UNITED STATES

Total money income	Male								Female							
	March 1969 CPS				March 1968 CPS				March 1969 CPS				March 1968 CPS			
	After allo-cation	Before allo-cation	Non-re-spond-ents	NA rate	After allo-cation	Before allo-cation	Non-re-spond-ents	NA rate	After allo-cation	Before allo-cation	Non-re-spond-ents	NA rate	After allo-cation	Before allo-cation	Non-re-spond-ents	NA rate
Number of persons..thousands.	67,611	57,460	10,151	15.0	66,519	58,042	8,477	12.7	74,889	65,786	9,103	12.2	73,584	66,440	7,144	9.7
Total....................	100.0	100.0	100.0	(x)	100.0	100.0	100.0	(x)	100.0	100.0	100.0	(x)	100.0	100.0	100.0	(x)
With income..................	92.4	91.8	95.9	15.6	92.4	91.9	95.8	13.2	64.8	62.9	79.0	14.8	63.8	62.0	80.2	12.2
Without income...............	7.6	8.2	4.1	8.1	7.6	8.1	4.2	7.0	35.2	37.1	21.0	7.3	36.2	38.0	19.8	5.3
Total with income.......	100.0	100.0	100.0	(x)	100.0	100.0	100.0	(x)	100.0	100.0	100.0	(x)	100.0	100.0	100.0	(x)
$1 to $999 or loss...........	11.5	11.7	10.3	14.0	12.3	12.6	10.2	11.0	30.8	31.6	26.1	12.6	33.9	34.6	28.9	10.4
$1,000 to $1,999.............	8.8	8.9	8.6	15.1	9.5	9.5	10.0	13.9	18.9	18.8	19.3	15.1	18.7	18.5	20.1	13.2
$2,000 to $2,999.............	7.3	7.3	7.4	15.8	7.8	7.9	7.3	12.3	12.1	11.9	13.2	16.1	12.4	12.4	12.0	11.8
$3,000 to $3,999.............	7.2	7.2	7.2	15.6	7.6	7.5	8.3	14.4	12.0	12.0	12.6	15.5	11.7	11.4	14.1	14.7
$4,000 to $4,999.............	7.1	7.0	7.3	16.0	7.6	7.6	7.3	12.8	8.7	8.7	8.9	15.1	8.2	8.2	7.9	11.7
$5,000 to $5,999.............	8.2	8.2	8.2	15.5	9.1	9.2	8.5	12.3	6.4	6.4	6.4	14.9	6.0	6.0	6.4	12.9
$6,000 to $6,999.............	8.6	8.7	8.2	14.7	9.5	9.5	9.1	12.7	4.4	4.4	4.6	15.5	3.8	3.7	3.9	12.6
$7,000 to $9,999.............	21.6	21.9	19.4	14.0	20.4	20.8	17.8	11.6	4.9	4.7	6.3	18.9	3.6	3.5	4.1	14.2
$10,000 to $14,999...........	13.7	13.5	14.6	16.6	11.0	10.8	12.5	15.0	1.4	1.3	1.9	20.7	1.0	0.9	1.7	20.9
$15,000 to $24,999...........	4.5	4.2	6.2	21.3	3.9	3.5	6.4	21.5	0.3	0.2	0.5	27.5	0.6	0.6	0.7	14.8
$25,000 and over.............	1.5	1.2	2.6	28.1	1.3	1.1	2.6	26.2	0.1	0.1	0.2	33.6	0.2	0.2	0.2	11.3
Median income................	$5,981	$5,955	$6,121	(x)	$5,572	$5,538	$5,814	(x)	$2,023	$1,976	$2,345	(x)	$1,859	$1,831	$2,078	(x)

x - Not applicable

Table C.--NUMBER OF PERSONS, NONRESPONSE RATE ON INCOME, AND MEDIAN INCOME IN 1966 OF PERSONS 14 YEARS OLD AND OVER, BY SEX*
(Numbers of persons as of March 1967)

Selected characteristics	Male			Female		
	Total (thousands)	Non-response rate	Median 1/	Total (thousands)	Non-response rate	Median 1/
Total.................	65,335	15.2	$5,306	72,224	9.3	$1,638
COLOR						
White.........................	58,501	15.6	$5,592	64,310	9.5	$1,715
Negro and other................	6,834	11.3	3,097	7,914	7.3	1,305
RELATIONSHIP TO FAMILY HEAD						
In families....................	60,772	15.4	$5,440	64,419	8.1	$1,549
Head.........................	43,750	16.3	6,348	5,172	13.1	2,364
Married wife present.......	42,553	16.3	6,389	(X)	(X)	(X)
Other marital status.......	1,197	18.0	4,346	5,172	13.1	2,364
Wife of head................	(X)	(X)	(X)	42,553	6.9	1,857
Other relative of head.......	17,022	12.3	1,178	16,694	9.9	862
Unrelated individuals..........	4,563	14.8	3,447	7,805	18.3	1,921
WORK EXPERIENCE 2/						
Worked in 1966.................	53,016	15.9	$5,903	35,295	10.8	2,364
Worked at full-time jobs.....	45,769	16.2	6,389	24,107	11.7	3,160
50 to 52 weeks.............	35,677	16.5	6,955	13,753	12.9	4,026
49 weeks or less...........	10,092	15.0	3,370	10,354	10.1	1,614
Worked at part-time jobs.....	7,247	14.4	1,025	11,188	9.0	827
50 to 52 weeks.............	2,581	16.1	1,680	3,490	10.4	1,504
59 weeks or less...........	4,666	13.5	796	7,698	8.3	592
Did not work in 1966...........	11,317	11.9	1,590	36,929	7.8	916
OCCUPATION AND EMPLOYMENT STATUS						
Employed civilians.............	46,934	16.3	$6,287	26,620	12.0	$2,948
Professional, technical, and kindred workers.............	6,190	20.9	8,773	3,760	13.6	4,841
Self-employed 3/............	713	42.2	13,346	184	28.0	2,406
Salaried...................	5,477	18.1	8,542	3,576	12.9	4,944
Farmers and farm managers....	1,791	14.6	3,459	75	16.9	(B)
Managers, officials, proprietors excluding farm.........	6,187	27.9	8,677	1,108	19.1	3,924
Self-employed 3/...........	1,799	34.1	6,711	339	25.5	2,163
Salaried..................	4,388	25.3	9,384	769	16.2	4,431
Clerical and kindred workers.	3,391	16.7	6,069	8,671	12.9	3,632
Sales workers................	2,825	19.2	6,537	1,860	13.4	2,155
Craftsmen, foremen, and kindred workers................	9,305	13.3	6,911	271	10.7	3,485
Operatives and kindred workers	9,535	11.5	5,665	4,128	9.7	2,986
Private household workers....	60	(B)	(B)	2,114	8.7	738
Service workers excluding private household...........	3,445	14.1	4,134	4,189	10.5	1,880
Farm laborers and foremen 3/..	1,075	5.7	1,734	355	7.1	605
Laborers, excluding farm and mine.......................	3,130	12.0	3,520	89	13.7	2,705
Unemployed.....................	1,685	10.4	2,747	1,335	5.9	1,455
Not in labor force or in Armed Forces........................	16,716	12.5	1,526	44,269	7.8	865

* Based upon 1105 computer allocation procedures. (X) Not applicable. (B) Base less than 75,000.

1/ Median based on number of persons with income.

2/ Based on February-April 1967 surveys; excludes members of the Armed Forces. Consequently, male subtotals for work experience does not add to "Total" on first line which includes members of the Armed Forces living off post or on past with their families.

3/ Includes a small number of unpaid family workers.

Table D-1.--MALES 14 YEARS OLD AND OVER BY TYPE OF INCOME IN 1968 AND 1967 AND PERCENT ALLOCATED, FOR THE UNITED STATES

(Income information for 1967 and 1968 was collected in the March 1968 and 1969 CPS surveys, respectively)

Income of specified type	March 1969 CPS		March 1968 CPS	
	Total thousands	Percent allocated	Total thousands	Percent allocated
TOTAL MONEY INCOME				
Total	67,611	15.0	66,519	12.7
Without income	5,110	8.1	5,066	7.0
With income	62,501	15.6	61,454	13.2
WAGE OR SALARY INCOME				
Total	67,611	9.7	66,519	7.4
Without income	17,374	4.0	17,210	2.4
With income	50,237	11.7	49,309	9.2
NONFARM SELF-EMPLOYMENT INCOME				
Total	67,611	7.0	66,519	4.7
Without income	62,078	5.8	61,238	3.6
With income	5,533	20.5	5,281	17.5
FARM SELF-EMPLOYMENT INCOME				
Total	67,611	6.4	66,519	3.9
Without income	64,459	6.3	63,391	3.7
With income	3,152	9.4	3,128	8.7
INCOME OTHER THAN EARNINGS				
Total	67,611	11.4	66,519	8.7
Without income	38,338	8.3	41,243	7.4
With income	29,274	15.5	25,276	10.7

Table D-2.—FEMALES 14 YEARS OLD AND OVER BY TYPE OF INCOME IN 1968 AND 1967 AND PERCENT ALLOCATED, FOR THE UNITED STATES

(Income information for 1967 and 1968 was collected in the March 1968 and 1969 CPS surveys, respectively)

Income of specified type	March 1969 CPS		March 1968 CPS	
	Total thousands	Percent allocated	Total thousands	Percent allocated
TOTAL MONEY INCOME				
Total	74,889	12.2	73,584	9.7
Without income	26,345	7.3	26,657	5.3
With income	48,544	14.8	46,927	12.2
WAGE OR SALARY INCOME				
Total	74,889	6.4	73,584	4.9
Without income	39,953	1.4	39,751	0.4
With income	34,936	12.2	33,833	10.2
NONFARM SELF-EMPLOYMENT INCOME				
Total	74,889	4.1	73,584	2.4
Without income	73,054	3.6	71,802	2.0
With income	1,835	24.6	1,783	20.4
FARM SELF-EMPLOYMENT INCOME				
Total	74,889	3.9	73,584	2.1
Without income	74,503	3.8	73,133	2.0
With income	386	16.3	451	11.3
INCOME OTHER THAN EARNINGS				
Total	74,889	9.2	73,584	6.3
Without income	54,305	7.1	54,497	5.2
With income	20,584	14.5	19,087	9.4

Table E.--HOUSEHOLD HEADS WITH SOURCES OF INCOME NOT REPORTED IN 1967, FOR THE UNITED STATES

Missing income items	Number (Unit count)	Percent
Total in sample	780	100.0
Income completely blank	232	29.7
Income partially blank	548	70.3
One type of income only	350	44.9
Wage or salary income (38)	104	13.3
Nonfarm self-employment income (39)	34	4.4
Farm self-employment income (40)	8	1.0
Social Security (41a)	19	2.4
Interest and dividends (41b)	155	19.9
Public assistance (41c)	9	1.2
Unemployment compensation (41d)	8	1.0
Private pensions, etc. (41e)	13	1.7
Two types of income	119	15.3
38, 39	3	0.4
38, 41a	3	0.4
38, 41b	47	6.0
38, 41c	1	0.1
38, 41d	3	0.4
39, 40	2	0.3
39, 41b	20	2.6
40, 41b	4	0.5
41a, 41b	18	2.3
41a, 41c	3	0.4
41a, 41d	2	0.3
41a, 41e	2	0.3
41b, 41c	2	0.3
41b, 41d	2	0.3
41b, 41e	7	0.9
Three types of income	26	3.3
38, 39, 40	1	0.1
38, 39, 41b	5	0.6
38, 40, 41b	4	0.5
38, 41a, 41b	2	0.3
38, 41b, 41d	1	0.1
39, 40, 41b	2	0.3
39, 41b, 41e	1	0.1
40, 41a, 41b	1	0.1
40, 41b, 41d	1	0.1
40, 41b, 41e	1	0.1
41a, 41b, 41d	1	0.1
41a, 41b, 41e	5	0.6
41b, 41d, 41e	1	0.1
Four types of income	18	2.3
38, 39, 40, 41b	4	0.5
38, 39, 41a, 41b	2	0.3
38, 39, 41b, 41c	1	0.1
38, 41a, 41b, 41d	1	0.1
39, 40, 41b, 41e	1	0.1
39, 41b, 41d, 41c	1	0.1
41a, 41b, 41d, 41e	1	0.1
41b, 41c, 41d, 41e	7	0.9
Five to seven types of income	35	4.5
38, 41a, 41b, 41d, 41c	1	0.1
39, 41a, 41b, 41d, 41c	1	0.1
41a, 41b, 41c, 41d, 41e	14	1.8
38, 39, 40, 41a, 41b, 41c	1	0.1
38, 41a, 41b, 41c, 41d, 41e	8	1.0
39, 41a, 41b, 41c, 41d, 41e	6	0.8
39, 40, 41a, 41b, 41c, 41d, 41e	4	0.5

Source: Sample households from Current Population Survey (March 1968 CPS), Bureau of the Census.

Table F.--FAMILIES WITH ONE OR MORE INCOME ALLOCATIONS, TOTAL AND ALLOCATED AGGREGATE FAMILY INCOME IN 1967

(Weighted count)

Total family income	All families 1/ (thousands) (1)	Families with one or more allocations 1/ (thousands) (2)	Col. (2) / Col. (1) Percent (3)	Aggregate family income		Col. (5) / Col. (4) Percent (6)	Col. (6) / Col. (3) Percent (7)
				Total income (thousands) 1/ (4)	Allocated income (thousands) 2/ (5)		
Total.............	49,834	8,852	17.8	450,232,537	45,240,948	10.0	56.2
None...................	173	49	28.3	--	--	--	--
$1 - $2,999 or loss.....	5,970	950	15.9	10,734,060	879,700	8.2	51.6
$3,000 - $6,999.........	14,381	2,218	15.4	72,609,669	6,492,086	8.9	57.8
$7,000 - $9,999.........	12,067	1,884	15.6	100,421,574	8,911,320	8.9	57.1
$10,000 - $14,999.......	11,196	2,086	18.6	133,389,144	13,438,012	10.1	54.3
$15,000 - $24,999.......	4,818	1,260	26.2	87,393,702	11,268,180	12.9	49.2
$25,000 and over........	1,229	403	32.8	45,684,388	4,251,650	9.3	28.4

1/ Eight rotation groups - 1967 income data.
2/ Mean allocated family income per interval for rotation groups 1 and 2 multiplied by column (2).

COMPUTER METHOD TO PROCESS MISSING INCOME AND WORK EXPERIENCE INFORMATION IN THE CURRENT POPULATION SURVEY

Emmett F. Spiers and Joseph J. Knott*
U.S. Bureau of the Census

Introduction

Missing information in interviews, a common problem found in all sample surveys which depend on public cooperation, can be handled primarily in three ways. One method is to show in detail reported income information only and allow the user to make all adjustments necessary to take account of missing data. The second method is to inflate respondents' data to control totals assuming that the characteristics of respondents and nonrespondents are alike. The third method is to impute missing data for nonrespondents based on information obtained from respondents with social and economic characteristics similar to those of nonrespondents.

Starting with the March 1962 Current Population Survey (covering 1961 income), when a respondent does not answer one or more of the income items in the CPS, an imputed amount is given to the nonrespondent based on the income information reported by a respondent with similar demographic and economic characteristics, such as age, sex, family status, color, number of weeks worked, earnings, and major occupational group. The income amount(s) assigned to a nonrespondent are the income amount(s) stored from the last respondent who otherwise had the same characteristics.

The March CPS Supplement includes eight questions on a person's income during the preceding calendar year.1/ In the February and April CPS supplements, questions are asked on a person's work experience (number of weeks worked last year, class of worker, occupation and industry of longest job, etc.) in the preceding year.2/

Information collected in the February, March, and April CPS supplements provides the basis for a number of Census Bureau publications. The wide interest in and use of data covering income and work experience makes the problem of income and work experience nonresponses of particular interest.

This paper, divided into two parts, outlines the computer methodology involved in this allocation procedure. The first part of the paper describes in summary fashion the computerized "allocation" procedure used to impute missing income and work experience information using the UNIVAC 1107 computer. The second part is a set of flow charts describing the logical flow used in imputing missing income information.

Allocation Procedures

A discussion of the field collection and editing procedures is beyond the scope of this paper, but it is noted that limited editing of the data is done in the field before the information on the questionnaires is converted to magnetic tapes for computer processing. After initial computer editing and format conversion, the February-April work experience data are added to the March income record for further joint processing. The matching of the work experience data with the income data is done on the basis of unique household and person numbers. If a February or April record corresponding to the March record is not found the work experience data are allocated based on the March data including age, sex, earnings, and family relationship. It should be noted that all March data excepting income have been edited and any necessary allocations have been made.

The actual processing of missing income and work experience data may be thought of as five separate steps or functions: (1) Work experience allocation; (2) earnings-work experience consistency edit; (3) earnings allocation; (4) postearnings allocation edit; and (5) "other income" allocation. Each of these steps or functions is discussed below in some detail.

(1) Work Experience Allocation Procedures (Chart A)

Work experience data are allocated in two situations:

1. When there is a March record and no corresponding February or April record can be found (nonmatch).

2. Work experience information had been allocated before the match occurred in a procedure not using earnings data. In order to have all missing work experience data allocated using the same procedure, the matched work experience data, with certain items previously allocated, are rejected and reallocated in this procedure using earnings data.

(2) Earnings-Work Experience Consistency Edit (Chart B)

A primary purpose of the earnings-work experience consistency edit is to eliminate illogical combinations of class of worker and type of income. For example, an employee who worked for a private company last year should have had some wage or salary income. Therefore, a person who worked more than 13 weeks last year (the smallest available interval) is edited in one of two ways if he reports no earnings in the earnings type consistent with his class-of-worker status.3/ In the first way an inconsistent earnings type is exchanged with the "none" in the consistent earnings type. In the second way the consistent earnings type is made NA4/ so that the person will go through the earnings allocation procedure. The effect of these

* Comments by members of the Consumer Income Statistics Branch and Dr. Murray S. Weitzman, Assistant Division Chief, Economic Statistics Programs, Population Division, are gratefully acknowledged.

procedures is to force a person to have an earnings amount consistent with his class-of-worker status.

Secondly, this edit is used to check a person's response to the number-of-weeks-worked-last-year question against his earnings data to see if the two are consistent. If a respondent reported he did not work last year, logically he should have had no earnings. In the present method a respondent must have less than $300 in earnings if he reports he did not work last year. If a respondent reports he did not work last year but also reports $300 or more in earnings, his work experience and earnings are inconsistent and either his earnings or work experience should be changed. The earnings data are assumed to be "correct" at this stage of processing and the work experience data are rejected and allocated from the last respondent with similar characteristics.5/

Thirdly, the income-work experience edit identifies and treats persons who have no work experience data and also have incomplete earnings data (three "NA's" or a combination of "NA's" and "Nones"). The allocation procedure is designed to use known information to allocate missing or unknown information. Very little economic information is available for a person whose work experience information is missing or unusable and whose earnings data are missing or incomplete. These cases are assigned a temporary earnings amount or "none" in the first earnings type which is NA. Wages or salaries earnings are checked first, nonfarm self-employed earnings next, and then farm self-employed earnings. Fortunately, this combination of missing earnings and work experience data does not occur frequently.

(3) Earnings Allocation Procedure (Chart C)

Missing earnings data are allocated from the reported earnings data of the last person with similar socioeconomic characteristics. The socioeconomic characteristics used in the earnings allocation procedure are:

(a) Number of weeks worked last year;6/
(b) occupation of longest job last year;
(c) family relationship;
(d) sex;
(e) age;
(f) color (white or nonwhite); and
(g) class of worker.6/

These six characteristics are combined in various ways to divide the records into 226 mutually exclusive classes. Each time a person answering all three earnings items is encountered, his three earnings items are stored temporarily in the appropriate class until another fully reported person having the same characteristics is encountered. When the next fully reported person belonging to the same class is encountered, his earnings data are stored replacing the previously stored data. This process continues throughout the processing. If at any time a person having any missing earnings data is encountered, he is assigned the earnings value(s) of the last person stored in his class.

The division of the records into these 226 classes is done with two objectives in mind. First, each class should be fairly homogeneous, i.e., the variables used in stratifying persons should be correlated with earnings. Second, the number of persons passing through each cell should be large enough to avoid having many persons' earnings allocated from the same fully reported person. The danger is that an atypical fully reported person could bias the allocated group if more than a few NA's are allocated using the fully reported person's atypical earnings data.

(4) Post Earnings Allocation Consistency Edit (Chart D)

The purpose of the post earnings allocation consistency edit is very much like the earnings-work experience consistency edit. However, only persons who were allocated one or more types of earnings are affected by this edit (an allocation may be either a dollar amount or a None).

Inconsistencies are caused by not controlling on class of worker uniformly throughout the allocation matrix. For example, if the last fully reported person in a class is a wage and salary worker, a self-employed worker with an NA in nonfarm self-employed earnings is likely to be allocated a None in nonfarm self-employed earning.

(5) Other Income Allocation Procedure (Chart E)

NA's in the five types of "other income" are treated by a separate allocation procedure essentially similar to the earnings allocation procedure. Earnings is used as a characteristic in the other income allocation matrix since it is considered to be correlated with other income types. For example, a person with high earnings is more likely to have dividend income and no public assistance than a person with low earnings. Other characteristics covered are: Family relationship, sex, worker-nonworker status, age, and color (white or nonwhite). These five characteristics are grouped to form 286 mutually exclusive classes. Like the earnings allocation procedure, the five other income items of the last fully reported person are stored in his appropriate allocation matrix cell.

Unlike earnings nonresponses, other income nonresponses are divided into two types. The first type occurs when a person being interviewed fails to indicate not only the amount of a specified type of other income but also fails to indicate whether he received that type. The second type occurs when a person indicates that he received a particular type of other income but fails to report the amount. To accommodate these two types of other income nonresponses, two distinct allocation matrices of equal size were developed, designated as matrix A and B, respectively. In matrix A, each type of other income is stored, either dollar values or "nones." On the other hand only dollar values are stored in matrix B. Therefore, when a nonrespondent indicates that he had received a specified source of other income but does not report the amount, he is allocated from matrix B, containing only dollar values. If, on the other hand a person refused to give any information (whether received or not and the amount) on a particular type of other income, that type is allocated from matrix A.

Footnotes

1/ In the March 1969 CPS the eight sources were: Earnings—(1) Money wages or salary; (2) net income from nonfarm self-employment; (3) net income from farm self-employment; Other income—(4) Social Security; (5) dividends, interest (on savings or bonds), income from estates or trusts or net rental income; (6) public assistance or welfare payments; (7) unemployment compensation, government employee pensions, or veterans' payments; (8) private pensions, annuities, alimony, regular contributions from persons not living in this household, royalties, and other periodic income.

2/ Because of the CPS sample design it is necessary to ask the work experience questions for one-quarter of the sample in the April CPS sample in order to have work experience data for all the households in the March CPS survey.

3/ The basic assumption in this edit is that a person who reports he worked more than 13 weeks last year should have some earnings of the type consistent with his class-or-worker status. Employees of incorporated private companies or any governmental unit must have a dollar value in wages or salaries to have his type of earnings and class of worker consistent. Self-employed farmers must have a dollar value in farm self-employment earnings to be consistent.

4/ "NA" means the information was not reported during the interview or was not available for some reason.

5/ The socioeconomic characteristics used in the work experience allocation procedure are: (1) Type of earnings; (2) amount of earnings; (3) age; (4) color (white or nonwhite); (5) sex; and (6) family relationship.

6/ Number of weeks worked and class of worker are not used uniformly throughout the matrix.

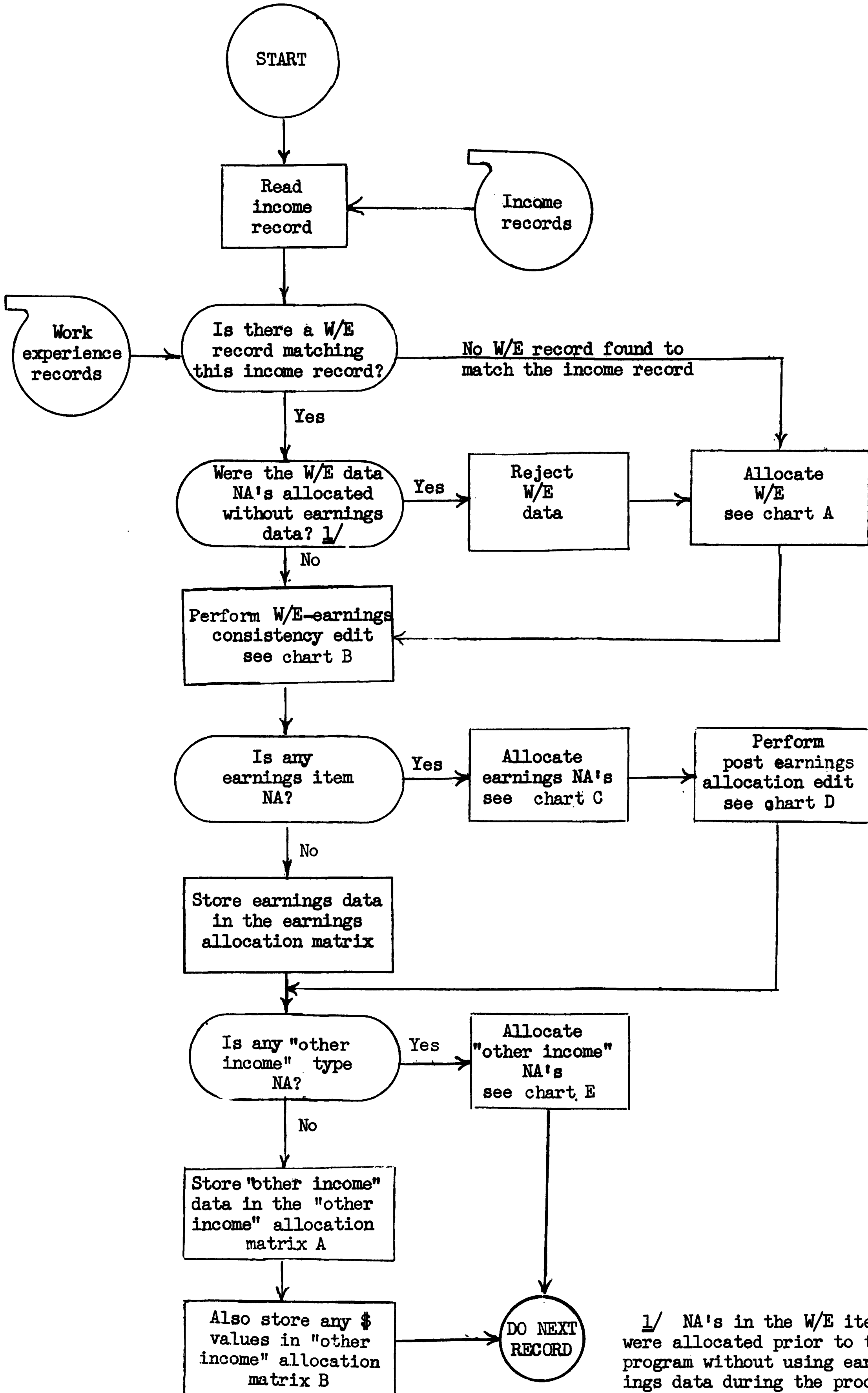
FLOWCHARTS OF THE INCOME ALLOCATION PROCEDURE

The following flowcharts outline the general methodology used to allocate missing income data collected in March of 1969 as a supplement to the Current Population Survey (CPS). The flow charts cover the six functional steps, each on a separate page. The first chart shows the overall flow of operations necessary for processing each income record. The other five charts (charts A to E) show the following: Chart A—Work Experience Allocation Procedures, Chart B—Work Experience-Earnings Consistency Edit, Chart C—Earnings Allocation Procedures, Chart D—Post Earnings Allocation Edit, and Chart E—"Other Income" Allocation Procedures.

Notations used in the flow charts are outlined below:

Notation	Representation
W/S	Money wages and/or salaries
NFSE	Net income from nonfarm self-employment
FSE	Net income from farm self-employment
Source A	Social Security payments
Source B	Dividends, interest (on savings or bonds), income from estates or trusts or net rental income
Source C	Public assistance or welfare payments
Source D	Unemployment compensation, government employee pensions, or veterans' payments
Source E	Private pensions, annuities, alimony, regular contributions from persons not living in the household, royalties, and other periodic income
Earnings	Earnings data—W/S, NFSE, and FSE
"Other Income"	Other income data—Sources A, B, C, D, and E
NA	Not available—any item of information which is not available during processing
W/E	Work experience data—number of weeks worked, class of worker, occupation of longest job, etc.
Matrix A	"Other Income" allocation matrix A contains the five "other income" items reported by the last preceding respondent in each class.
Matrix B	"Other Income" allocation matrix B contains, for each type of income within each class, the dollar value reported by the last preceding respondent having a dollar value in that particular type of income.

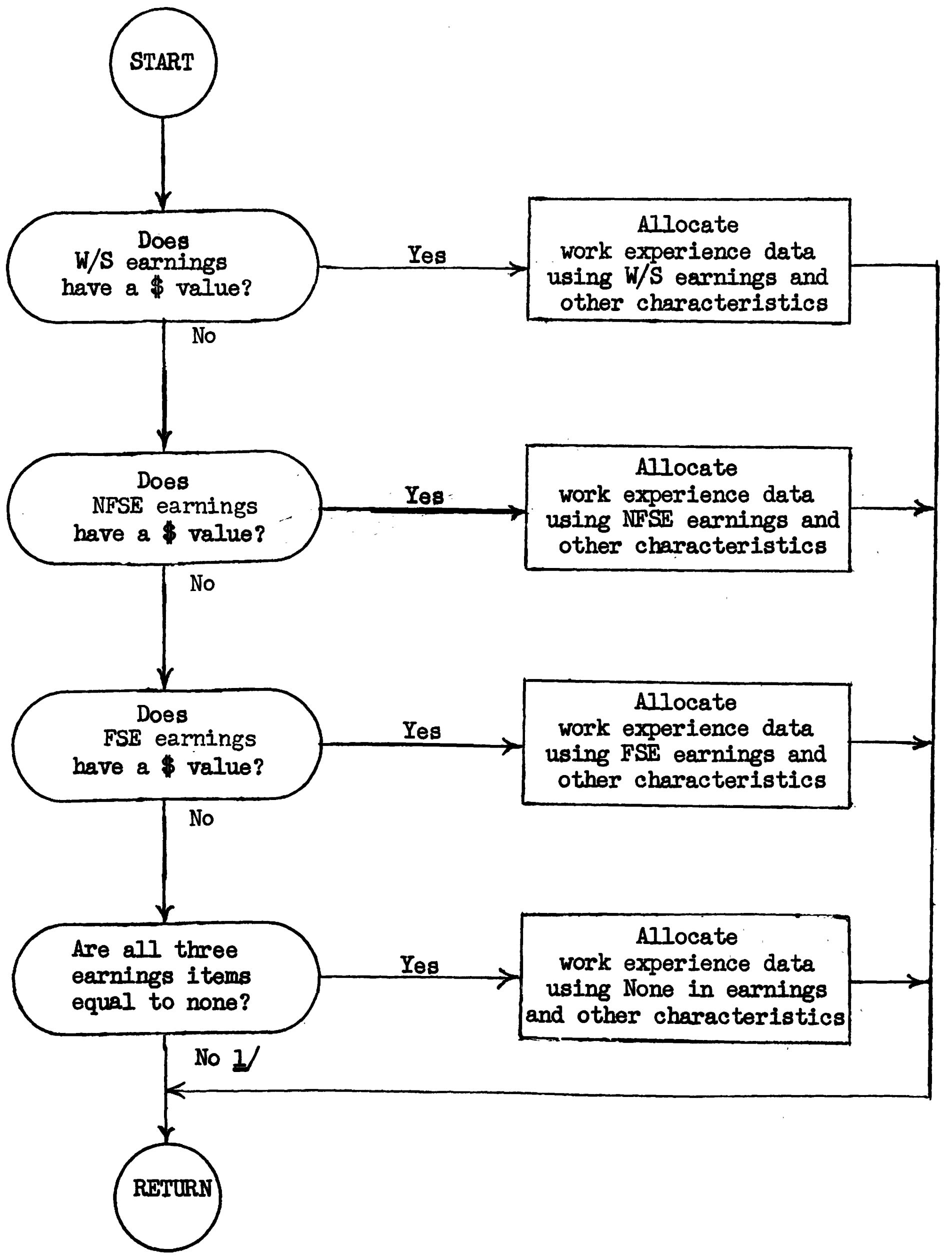
INCOME ALLOCATION PROCEDURE
(Main Flowchart)



1/ NA's in the W/E items were allocated prior to this program without using earnings data during the processing of the February CPS.

Work Experience Allocation Procedure

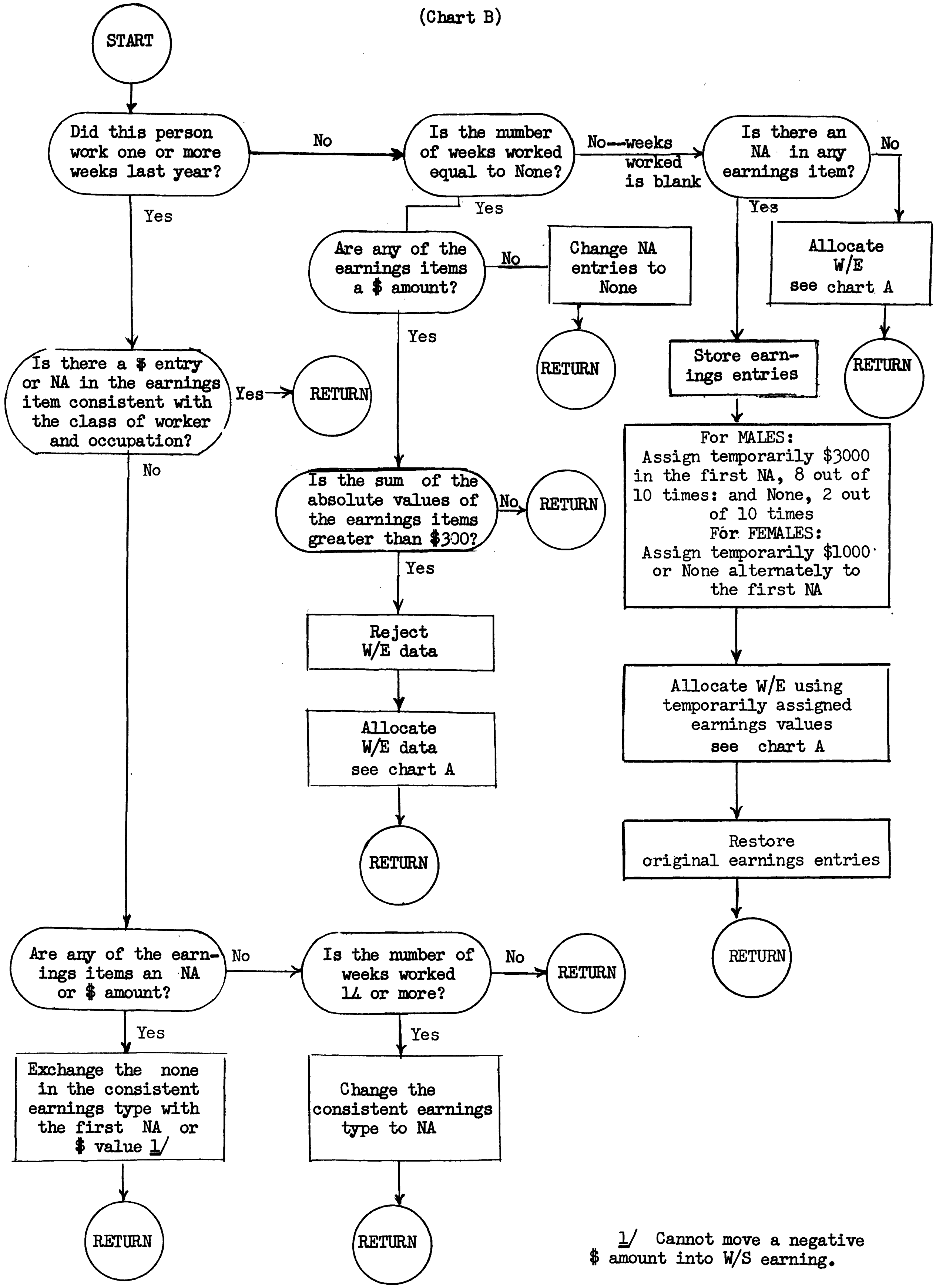
(Chart A)



1/ This path is a small group of people who did not report work experience and have NA's in earnings or a combination of NA's and none's.

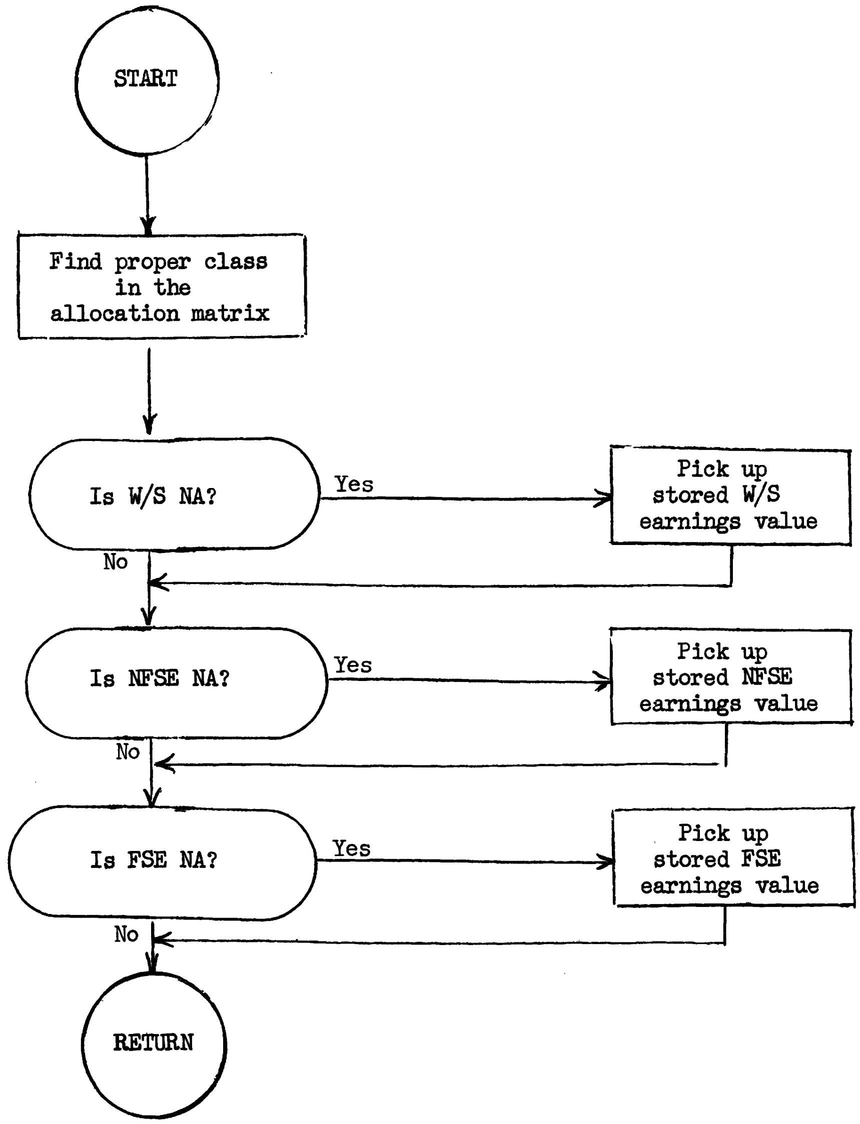
Work Experience-Earnings Consistency Edit

(Chart B)

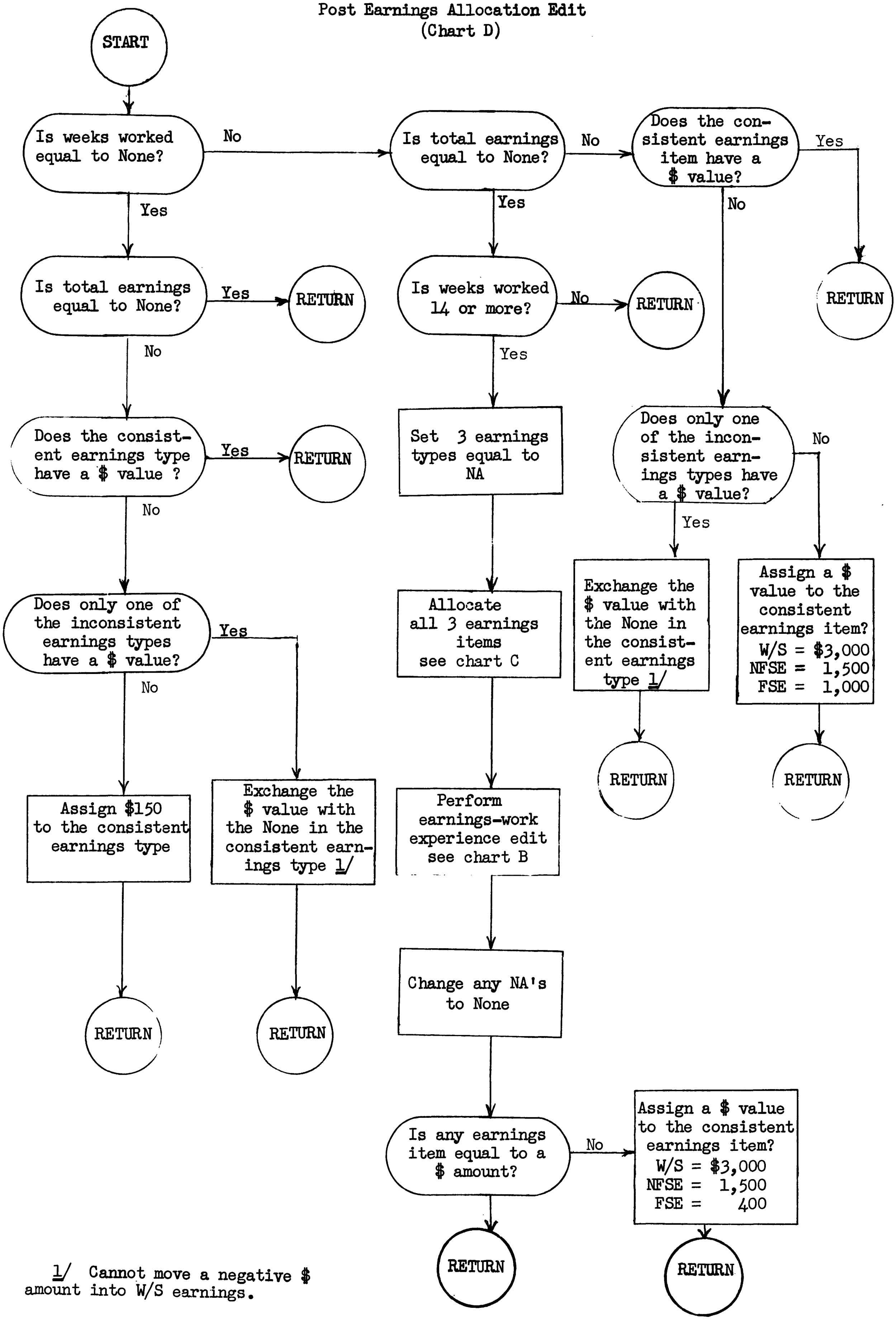


1/ Cannot move a negative $ amount into W/S earning.

Earnings Allocation Procedure
(Chart C)



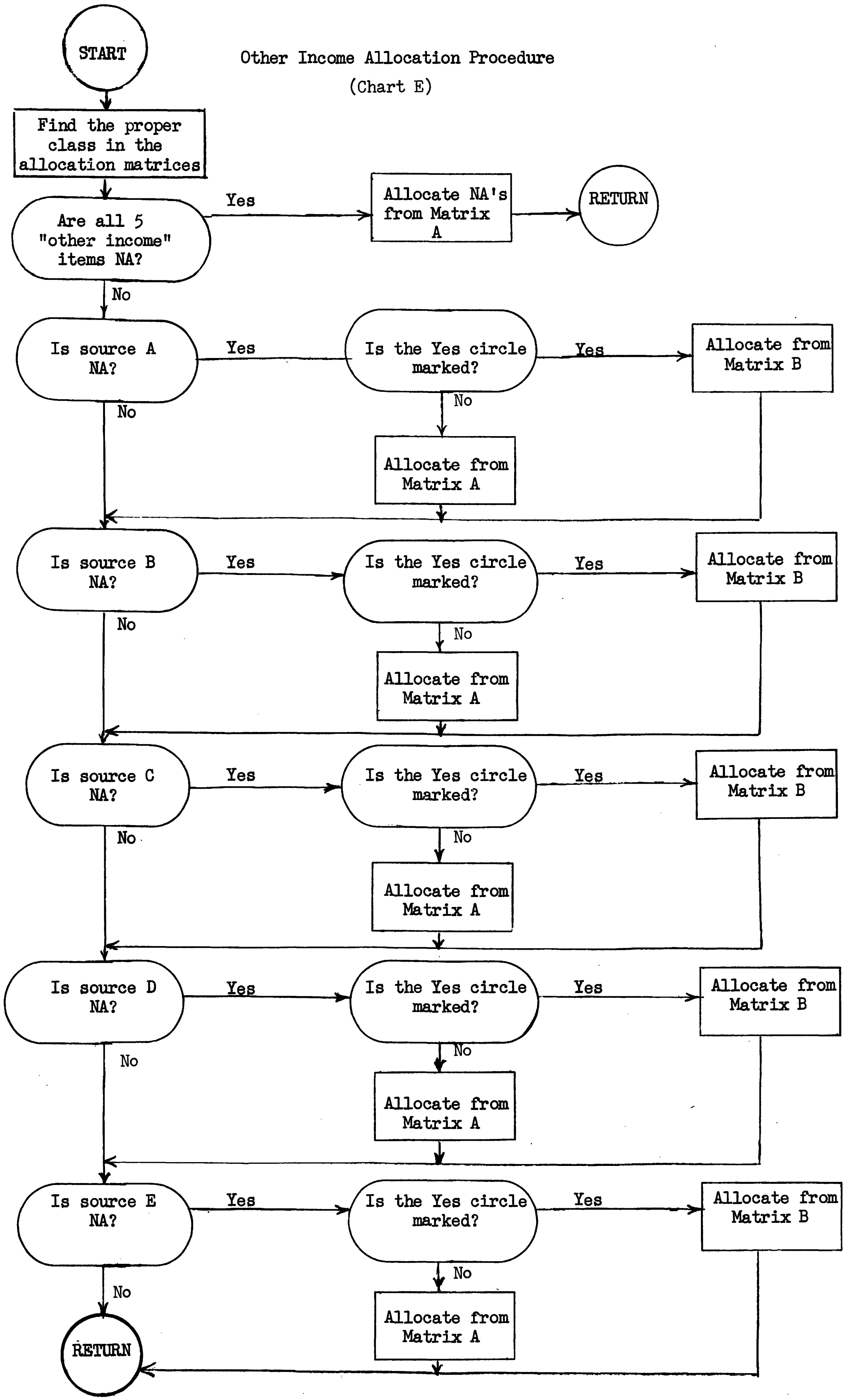
Post Earnings Allocation Edit
(Chart D)



1/ Cannot move a negative $ amount into W/S earnings.

Other Income Allocation Procedure

(Chart E)



A GRAPHIC TECHNIQUE FOR PROJECTING FAMILY INCOME SIZE DISTRIBUTION

Mitsuo Ono*
U.S. Bureau of the Census

Introduction

Projections of family income distribution are used widely not only by businessmen in developing marketing plans but also by government officials in formulating public programs, e.g., urban development studies. These projections can be calculated in a variety of ways, depending upon the availability of resources, e.g., time, money, skilled manpower, equipment. However, all of these projections, whether simple or complex, have one thing in common: the adequacy of the projections depends ultimately on the validity of the assumptions used to make the projections, and usually these assumptions are no more than calculated historical guesses. Thus, under certain circumstances, the use of relatively simple estimation methods may yield results comparable to that obtained from use of more complicated methods, e.g., regression equations, because use of both of these methods encompasses the same growth factors.

The primary purpose of this paper is to describe a relatively simple graphical technique using logarithmic normal probability (lognormal) graph paper by which projections of family income size distribution can be developed. The discussion is divided into four parts: (1) An outline of the graphical method; (2) an outline of an arithmetical method; (3) comparison of results obtained from using other estimation procedures; and (4) an example of the use of the graphical method.

The Graphical Method

The technique involves essentially the following: (1) Plot historical data to determine the relative "stability" of the cumulative income distribution; (2) if the distribution "shape" appears to be similar historically, obtain the median family income value for the most current distribution available; (3) compute a projected median value, e.g., using historical growth rates of median family income; (4) arithmetically extrapolate the benchmark year income levels to projected year income levels (at given cumulative percentage levels); and (5) plot "new" distribution on graph paper and obtain required information. A specific example of this procedure is outlined below. Lognormal graph paper is used because it tends to reduce the skewness of the income distribution curve. That is, some skewed distributions are transformed into normal probability distributions by taking the logarithm of the variates. If the distribution is lognormal, the ogive generally forms a straight line when it is plotted on lognormal graph paper (in which the ordinates are scaled logarithmically and the abscissa scaled in accordance with the normal distribution function). The slope of the line is related to the standard deviation of the logarithm of the variate. In a lognormal distribution, it is assumed that the magnitude of a variate changes at a rate proportional to the value of the previous variate. (See publication entitled The Lognormal Distribution by J. Aitchison and J.A.C. Brown, Cambridge University, 1957, for further details.) There are other types of graph papers, e.g., arithmetical, log-log, probability, and semi-log, which can be used just as well for the graphic method but it appears that the lognormal graph paper is more convenient to work with because the linearity of the plotted distributions generally extends over a wider range on this type of graph paper than others.

Assume a need exists for a projected (1985) family income distribution expressed in 1967 dollars. First, we plot the cumulative family income distribution (in 1967 dollars) for 1952, 1957, 1962, and 1967 to examine the relative stability of the income distributions over the past 15-year period. We find that the "shapes" of the income distributions are fairly constant (see attachment 1). We then compute the rate of increase of median family income between 1957 and 1967. We then assume that the growth rate of median family incomes between 1957 and 1967 also applies for the period between 1967 and 1985 (in constant dollars). Thus: the 1957-1967 average annual rate comes up to about 3.1 percent per annum (for 10 years). We then apply this growth rate to 18 years, resulting in a total (compounded) increase of 1.73 times the median income in 1967, i.e.,

$$\frac{\text{1967 Median income \$7,974}}{\text{1957 Median income \$5,889}} = \text{1.354 or about 3.1\% compounded annually}$$

$$(1.031)^{18} = 1.732$$

* The views expressed here are not necessarily those of the Bureau of the Census.

The next step is to multiply this net increase against the 1967 income levels to obtain the 1985 income levels (in 1967 dollars). Thus:

1967 Cumulative Percentage of Families	1967 Income Intervals	1985 Income Intervals
2.1	Under $1,000	Under $1,730 (1,000 X 1.73)
6.5	Under $2,000	Under $3,460 (2,000 X 1.73)
12.5	Under $3,000	Under $5,190 (3,000 X 1.73)
25.3	Under $5,000	Under $8,650 (5,000 X 1.73)
41.4	Under $7,000	Under $12,110 (7,000 X 1.73)
50.3	Under $8,000	Under $13,840 (8,000 X 1.73)
58.6	Under $9,000	Under $15,570 (9,000 X 1.73)
65.7	Under $10,000	Under $17,300 (10,000 X 1.73)
88.1	Under $15,000	Under $25,950 (15,000 X 1.73)

This procedure, of course, assumes that the income distribution curve has shifted over time with insignificant changes in the slope or "inequality" of the relative distribution. We plot on the lognormal graph paper the projected 1985 income levels (at the 1967 cumulative percentage points) for various income levels (see attachment 1). If we want a more refined fit, we would use more detailed income intervals which would, of course, take into account more of the "curvature" of the cumulative income distribution. The final step is to connect the points graphically and read off the required information. Thus using this method, it is estimated that in 1985 about 9 percent of all family units would have incomes of less than $4,000 (in 1967 dollars). That is, the $4,000 line cuts the projected 1985 income distribution at about the 9 percent point (found at the bottom of attachment 1).

The Arithmetical Method

The same results obtained graphically can be also calculated as shown below. The first step is to calculate the total net increase in median family incomes between the benchmark year and the projected year. We found this increase to be 1.73 between 1967 and 1985 (in constant 1967 dollars).

COMPUTATIONS FOR ARITHMETICAL METHOD

1967 Income levels	1985 Income levels (1967 values X 1.73)	Cumulative percentage of families in 1967	1985 Income levels	1985 Cumulative percentage of families using arithmetical method	1985 Cumulative percentage of families using graphic method
Under $1,000	$1,730	2.1	Under $1,000	$\frac{\$1,000}{\$1,730}$ (2.1)=(.58)(2.1)=1.2	NA
Under $2,000	$3,460	6.5	Under $2,000	$\frac{\$2,000}{\$1,730}$ (2.1)=(1.16)(2.1)=2.4	2.8
Under $3,000	$5,190	12.5	Under $3,000	$\frac{\$3,000}{\$3,460}$ (6.5)=(.87)(6.5)=5.7	5.8
Under $5,000	$8,650	25.3	Under $5,000	$\frac{\$5,000}{\$5,190}$ (12.5)=(.96)(12.5)=12.0	12.0
Under $7,000	$12,110	41.4	Under $7,000	$\frac{\$7,000}{\$8,650}$ (25.3)=(.81)(25.3)=20.5	19.0
Under $8,000	$13,840	50.3	Under $8,000	$\frac{\$8,000}{\$8,650}$ (25.3)=(.92)(25.3)-23.3	23.0
Under $9,000	$15,570	58.6	Under $9,000	$\frac{\$9,000}{\$8,650}$ (25.3)=(1.04)(25.3)=26.3	26.0
Under $10,000	$17,300	65.7	Under $10,000	$\frac{\$10,000}{\$8,650}$ (25.3)=(1.16)(25.3)=29.3	30.0
Under $15,000	$25,950	88.1	Under $15,000	$\frac{\$15,000}{\$15,570}$ (58.6)=(.96)(58.6)=56.3	56.0

NA Not applicable.

The second step is to multiply this rate against the income levels. Thus, we multiply $1,000 by 1.73 to obtain $1,730, etc., as we had done before in the graphic method. The third step is to estimate the cumulative percentage levels in 1985 for the original income levels in 1967. We can also use either extrapolations or interpolations in order to take account of the "curvature" of the income distribution.1/ This adjustment is needed if the income intervals that are being used in the calculation are large. The following table shows the computational method and a comparison of the cumulative percentage rate obtained from both the arithmetical method and the graphic method.

In the computations, six income levels, e.g., the "$5,000 and under" level in 1985 were interpolated while three income levels were extrapolated, e.g., the "$9,000 and under" level in 1985. Overall, the results obtained from using the graphic method and the arithmetical method are similar and either method can be used depending on one's preference.

A Comparison of Projections

Projections of families receiving under $3,000 income were developed for 1975 (in constant 1965 dollars) using two methods:

1. A regression equation to estimate the change in median family income from 1965 to 1975. Thus,

$$Y_F = \underset{(32.62)}{-1,172} + \underset{(0.008)}{0.749}\,(X_{PF}) \quad N = 16$$

where Y_F = Median family income

X_{PF} = Personal income per family

Personal income in 1975 was estimated from GNP projections described in Joint Committee Print, U.S. Economic Growth to 1975: Potential and Problems, 89th Congress, Second Session, U.S. Government Printing Office, Washington, 1966. GNP projections and implicit GNP price indices are shown on page 16 of this reference. Model B (assumes a 4.5 percent annual rate of growth for real GNP, 1966-1975) was used. This model assumed a 5.5 percent average personal savings rate for the period 1961-1965 and a 4 percent unemployment rate (see page 9 of cited reference). The 1975 GNP (in 1965 dollars) was converted to estimated Personal Income by using the PI/GNP ratio shown in current dollars in the cited reference. The projected PI value was divided by the projected number of families to obtain PI per family in 1975. The total increase in median family income using this regression procedure was 1.37 times the median family income in 1965. Using the graphic method, the percentage of families receiving under $3,000 (in 1965 dollars) was estimated at about 10.0 percent in 1975.

2. The average annual rate of increase in median family income between 1955 and 1965 (in 1965 dollars) was calculated to be about 2.9 percent or a total increase in median family income of 1.33 over a 10-year period. This procedure resulted in an estimate of about 11 percent of families receiving under $3,000 (in 1965 dollars). It is interesting to note that these percentage rates are fairly comparable to those derived from using regression methods. For 1975, the following percentage of families receiving under $3,000 was computed:

Reference source	Assumption	Percentage of all families under $3,000	Basic relationship
"Poverty by Color and Residence--Projections to 1975 and 1980," by J. Patrick Madden, paper presented at American Agricultural Economic Association Meeting, Bozeman, Montana	4% Unemployment rate and 1959-1966 data fitted to log-log equation (1964 constant dollars)	10.7	Percentage of families with $3,000 or less is related to median income and unemployment rate

The point of this discussion is that under certain assumptions, relatively simple extrapolation methods may be used to derive approximations which may be almost similar to results derived from use of more complicated methods.

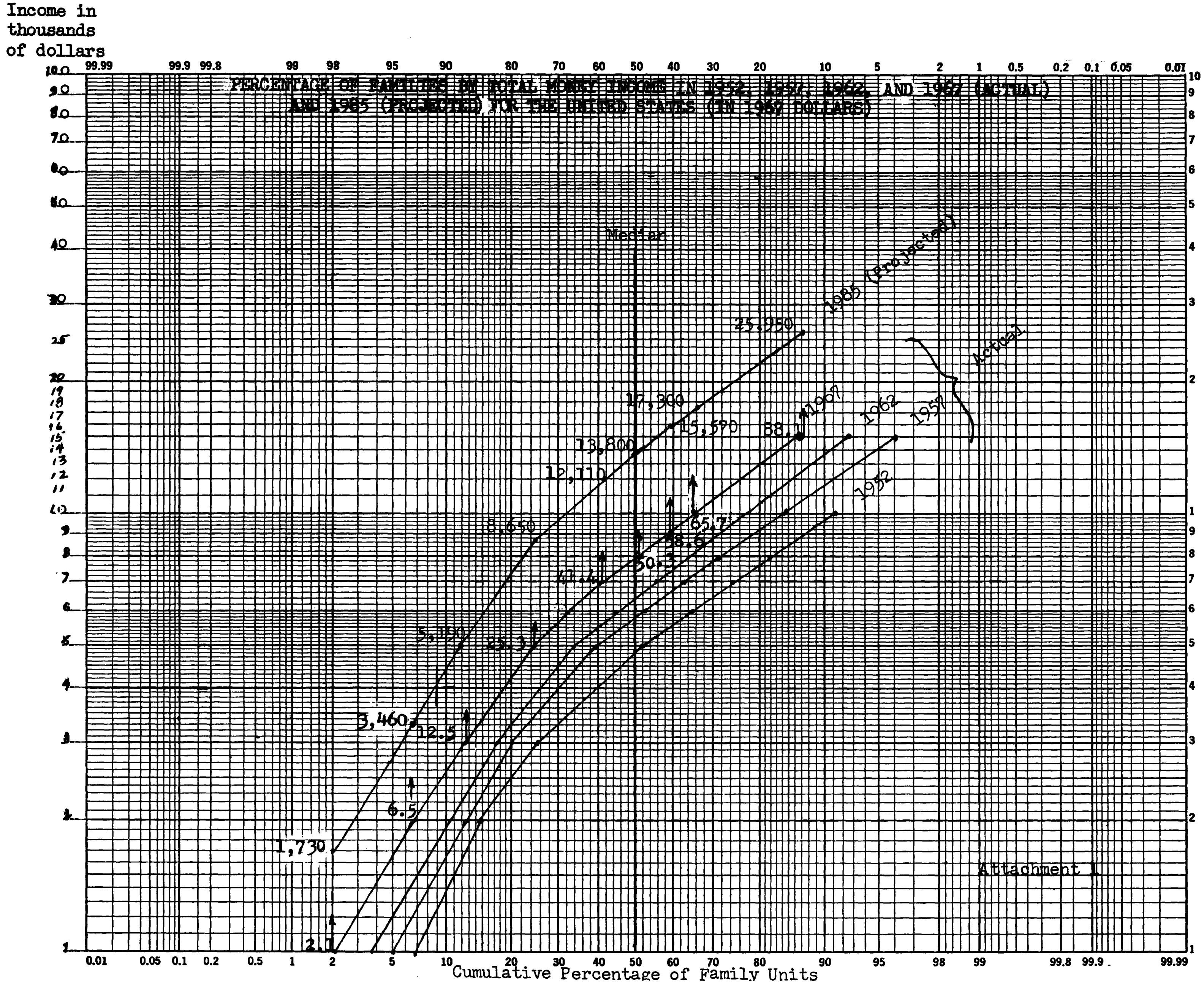
An Example of Use of the Graphical Method

Advantages of the graphical method are the simplicity and flexibility introduced in obtaining the desired information. Thus, we may find that historically the "curvature" of the income distribution has changed over time. Instead of taking a constant percentage change for each income level value, it is possible to apply differential rates of increase to different income levels which would take into consideration the change in relative "inequality" of the distribution over time. Instead of keeping the "slope" constant, changes in both the "slope" and the median income level are taken into account. The "slope" related to the standard deviation in turn is related to measures of inequality, e.g., the Gibrat inequality measure.

As a specific example of this technique, we may relate historical changes in summary measures of inequality to percentage changes in income levels, e.g., quintile values. In turn, these changes in values can be related to changes in socioeconomic characteristics of that group. In this way, relationships are obtained for quintile groups, associating changes in summary measures of inequality with changes in social and economic characteristics of different income groupings.

FOOTNOTE

1/ This comment, attributed to Mr. Albert Mindlin, Chief Statistician, Government of the District of Columbia, is gratefully acknowledged.


PERCENTAGE OF FAMILIES BY TOTAL MONEY INCOME IN 1952, 1957, 1962, AND 1967 (ACTUAL) AND 1985 (PROJECTED) FOR THE UNITED STATES (IN 1967 DOLLARS)
Income in thousands of dollars
Cumulative Percentage of Family Units
Median
1985 (Projected)
Actual
1967
1962
1957
1952
25,950
17,300
15,570
13,800
12,110
8,650
5,190
3,460
1,730
88.1
65.7
58.6
50.3
41.4
25.3
12.5
6.5
2.1
Attachment 1

YEARLY CHANGES IN HOUSEHOLD COMPOSITION AND FAMILY INCOME

Marshall L. Turner, Jr., Bureau of the Census

INTRODUCTION

Economists, poverty analysts, and demographers are interested in how households change in composition and the effects of such changes on the family's economic situation. From the viewpoint of the economist, the problem of changes in household composition centers on how the structure and size of family income varies as members enter or leave the unit. From a similar perspective, the poverty analyst is interested in knowing if changes in household membership may move the family in or out of poverty. The demographer's main reason for studying gross changes in household composition is to observe the frequency of such changes and assess their effects on the growth and structure of the population and family unit.

This need to investigate problems in family membership dynamics has been expressed by several writers. Martin has pointed out that apparent year-to-year changes in statistics on family and individual households in poverty may be misleading unless one can obtain comparable data on the gross compositional changes these households might have undergone during the same time period.[1] Fisher similarly notes that the consumer behavior of families varies with changes in family personnel and that such phenomena need to be studied by means of a panel survey.[2] Miller, too, calls attention to the need for longitudinal data on family composition and income for seeking answers about the incidence and nature of poverty.[3] Given these expressed needs for statistics on gross changes in family composition and income the purpose of this paper is to describe preliminary tabulations of such longitudinal data from the 1967 Survey of Economic Opportunity conducted by the Bureau of the Census.

Survey of Economic Opportunity.--Under the sponsorship of the Office of Economic Opportunity, the Bureau of the Census conducted two surveys of economic opportunity in February-March 1966 and February-March 1967. These surveys offer two sources of longitudinal data. First, there is the portion of the national random sample of households (family and primary individual units) that was interviewed in 1966 and reinterviewed in 1967. To utilize this longitudinal potential of the overlapping sample, the computer tape files containing the basic data for the uniquely serialized households are now being collated in a computer-record matching operation.

Although the matching work is still incomplete, the 1967 SEO file alone provides a second source of longitudinal data on changes in household composition. In the 1967 SEO, enumerators revisited each physical address included in the 1966 survey and noted any changes in the interview units' members since the first interview. For persons that had entered or left the unit, information on the length of tenure in the unit, personal income and other characteristics were recorded. If the enumerator found that none of the 1966 household members for a given interview unit were living at the same address in 1967, the fact was noted, and any new family at that address was interviewed. With this type of procedure, two types of household interview units were included in the 1967 SEO. First, there were units that were interviewed for the first time in the SEO, and second, there were households interviewed in both the 1966 and 1967 surveys. When the questionnaire information was transcribed onto computer tape records, each household head's record was given a code which indicated whether this interview unit had been enumerated in both 1966 and 1967. In preparing the current longitudinal tabulations, a computer program was written to search the 1967 SEO tape file and select for tabulation those household units that were coded as having been interviewed in both 1966 and 1967. This group of units makes up the households referred to as "matched" in this paper.

MATCHED HOUSEHOLDS

Of the 61.3 million households represented in the 1967 SEO, approximately 75 percent (45.9 million) were also interviewed in the 1966 survey and were classified as matched households. This statistic of 75 percent agrees closely with the corresponding number of 78 percent for households matched in an earlier, but similar, study involving units interviewed in both the March 1964 and March 1965 Current Population Surveys (CPS).[4] From a further comparison between these SEO longitudinal data and the earlier CPS matching study, it is observed that 90.1 percent of the SEO matched households had white heads as compared with 90.2 percent of the CPS matched households. In the CPS operation, the characteristics of the matched and nonmatched households indicated that the nonmatched households were due mainly to entire households that moved from or into a sample address during the one-year period and were interviewed in only one of the two years.[5] Based on the similarity of the above statistical comparisons between the SEO and CPS matched household data and the similarity of the methodology and concepts used in these surveys, it seems reasonable to assume that the households not matched in the present SEO tabulation were also due to mobile households as was the case in the previous CPS study. It should be pointed out, however, that until data become available on the characteristics of the nonmatched SEO households, this inference about the nature of these nonmatches is indirect.

RESULTS

Characteristics of households with changes.--As shown in Table 1., approximately 17.2 percent (7.9 million) of the 45.9 million matched SEO households underwent a change in membership between 1966 and 1967.[6] Only approximately 17 percent of the white households had a change in membership as compared to 24 percent of all nonwhite households. With respect to poverty status of the household, the Table 1 data show only slight differences in household composition changes. Approximately 17.4 percent of the families above the poverty line lost and/or gained members as compared with 17.6 of the poor families.

Types of changes.--The data in Table 2 classify matched households that underwent a change in composition by the number of members who entered or left the household, the color of the household head, and the relative poverty status of the unit. Looking first at the types of gross changes, it is observed that approximately 5.4 million of these changed households lost members between 1966 and 1967. These households represent approximately 69 percent of all the matched households that had a membership change. Fifty-seven percent of these changing households lost only one person; 8.4 percent lost two people; and approximately 3.5 percent lost three or more members.

In contrast to these households that only lost members approximately 2.5 million (31-percent) of the matched households with changes in composition had members join. Roughly 28-percent of the changing units added only one member and the remaining 2.6 percent added two or more members. Although net change data are not available for these households that added members, approximately 393,000 (see Table 2a.) of the units lost and gained at least one member. Therefore, it is possible that for this group of households, an added member was offset by a lost member, and the unit had no overall change in size.

Comparing these changed households according to color of head and poverty status reveals only a few differences. With respect to color, the Table 2. data indicate that households with white heads seemed slightly more likely (69.3-percent as compared to 64.5 percent) to lose a member than did households headed by nonwhites. From a complementary perspective, the nonwhite headed households appeared more likely to have had members join (35.5 percent as compared to 30.7 percent). Comparing these households in poverty in 1966 to the total population of households undergoing a change in composition indicates that nearly 80 percent of the poor households (as compared to 69 percent of all the changed households) had members leave in this one-year period.

Reasons for changes.--The statistics presented so far have been aimed primarily at describing the overall types of compositional changes undergone by the SEO matched households. Referring now to Table 3., it is observed that approximately 1.9 million children joined the 7.9 million matched households that underwent a membership change. Most of these children (78-percent or 1.5 million) who joined these households were born between the 1966 and 1967 interviews. The remaining 21 percent of these children joined for a variety of other reasons including such events as deaths in families, divorce of parents, and other unspecified reasons.

Table 3. also deals with the reasons for adults (persons 14 years old or over) joining these matched units. Of the 1.2 million joining adults, approximately 13 percent joined the new household in order to take a job. Of the other reasons cited for these joining adults, approximately 13 percent moved in to get married, and another 18 percent were returning from an institution.

White-nonwhite comparisons with respect to the reasons for persons joining the matched households revealed few significant differences. Births accounted for fewer of the children joining nonwhite households than for white households. Approximately 81 percent of the additional children in white households were reported to be accounted for by births as compared with only 61 percent of the joining children in households headed by nonwhites.

Table 3a. provides estimates of the reasons for persons leaving these matched households. Of the 5.3 million leavers who reported such reasons, approximately 15 percent left in order to get married; 32 percent left because they joined the Armed Forces; 25 percent entered institutions; eight percent died; and 16 percent left because of divorce. The remaining approximately four percent reported "other" reasons for leaving.

Overall, these data on reasons for changes in household membership indicate that births, marriages, divorces and deaths were the major reasons for changes in composition. Purely economic reasons for entering or leaving households ranked relatively low in frequency, but the more common change reasons cited above did have economic implications.

Income of leavers and joiners.--Table 4. presents data on the income of persons who left or joined the matched households between 1966 and 1967. Although these data alone do not show the direct relationship between changes in household composition and income, certain inferences can be made about the economic impact of these persons entering and departing from the household. Overall, there were approximately 6.6 million households that lost or gained approximately 8.2 million income-contributing persons. Distinguishing between the households that had joiners or leavers, an estimated 819,000 households had an estimated 878,000 joiners. On the other hand, approximately 7.3 million persons left an estimated 5.8 million households.

Analyzing the income of the joiners first, it is observed that roughly 43 percent of these persons reported no income. Another 21.4 percent had incomes that ranged between $1.00 and $1,499; 25 percent reported incomes in the $1,500 to $4,999 range, while approximately six percent had at least $5,000 personal income.

Among the persons that left the matched households during this year, only six percent were in the "loss or no income" category. Approximately 84 percent of these persons had personal incomes that ranged between $1.00 and $1,499. Of the remaining leavers, approximately six percent had incomes in the $1,500 to $2,999, and slightly over three percent had incomes in excess of $3,000. "Losing members" seemed to be a more economically meaningful type of compositional change than "adding members." In other words, families that lost members appear to have had a loss of income disproportionately greater than the income gain experienced by families that added members.

Income contribution to the family by joiners and leavers.--These longitudinal data also provide more direct evidence on the relationship between family income and the income contribution of persons who joined or left the matched households. The data in Table 5. indicate that in approximately 70 percent of 5.7 million matched households, leavers and joiners contributed less than one percent of the family income. When this overall statistic is broken down by level of family income, it is observed that the higher the family income, the smaller the proportion of households in which the transient members made no contribution to the family's income. Seventy-seven percent of the households with family incomes under $2,000 and joiners/or leavers who contributed less than one percent of the family income. For families with incomes in the $2,000 to $4,999 range, the percentage was 72 and for families with incomes of $5,000 or more, this number was approximately 70 percent.

It is also interesting to note the relationship between a significant level of income contribution with respect to the level of family income. The Table 5 data show that approximately 17 percent of the households with family incomes under $2,000 had part-year members who contributed at least 50 percent of the family income. Similarly, nearly 11 percent of the households in the $2,000 to $4,999 family income category had joiners or leavers who made income contributions of this magnitude. For the households with family incomes of $5,000 or more, only approximately 10 percent had part-year members who contributed at least 50 percent of the family's income.

Summary.--The intent of the work described in this paper is to demonstrate the types of longitudinal data that can be produced from surveys such as the Survey of Economic Opportunity and to illustrate how these statistics can be used to study such problems as the relationship of household composition changes to changes in the family's economic situation. Moreover, the results of this project confirm the earlier finding that approximately 17 percent of all matched households underwent a compositional change in the 12-month study period. In addition, these estimates indicate that households with white heads are less likely to undergo such membership changes than are households headed by nonwhites.

These data also show that most of the persons who joined these households did so for reasons that can be classified as strictly demographic changes. Approximately 57 percent of the children and adults who joined these units cited births, marriages, and deaths for the reasons they made these changes.

With regard to the economic impact on the family of these gross changes in household composition, these longitudinal data show that persons who left households are more likely to have been making a significant contribution to the family income than are persons who joined such units. Overall, in most households with leavers and/or joiners, 71 percent, part-year members contributed less than one percent of the family income. Furthermore, as family income increases, the proportion contributed by leavers and joiners declines.

REFERENCES

1) Martin, Margaret E., (1966). Memorandum to Raymond T. Bowman, "Program for information on changes in household composition," Bureau of the Budget.
2) Fisher, Janet A., "Family life cycle analysis in research on consumer behavior," Consumer Behavior, 2, 3.
3) Miller, Herman P., (1965). "Dimensions of poverty," The Poverty as a Public Issue, ed. Ben B. Seligman, 47.
4) Turner, Marshall L., Jr., (1967). "A new technique for measuring household changes," Demography, 4 (1), 344.
5) Ibid., 344-347.
6) Ibid., 344.

TABLE 1.--MATCHED HOUSEHOLDS BY GROSS CHANGES IN COMPOSITION BETWEEN 1966 AND 1967 BY COLOR OF HEAD AND MEMBERS ADDED OR LOST

(Numbers in Thousands)

Matched Households by Type of Change in Membership	Total Households		White Head		Nonwhite Head		Percentage of households	
	Number	Percent	Number	Percent	Number	Percent	Poor	Nonpoor
All matched households....................	45,878	100.0	41,351	100.0	4,527	100.0	100.0	100.0
No change in members....................	37,980	82.8	34,538	83.5	3,442	76.0	82.4	82.6
Change in members......................	7,898	17.2	6,813	16.5	1,085	24.0	17.6	17.4
Households with a change in members.......	7,898	100.0	6,813	100.0	1,085	100.0	N.R.	N.R.
Households that only added members......	2,082	26.4	1,785	26.2	297	27.4	N.R.	N.R.
Households that lost members.............	5,816	73.6	5,028	73.8	788	72.6	N.R.	N.R.
Households that only lost members.......	5,423	68.7	4,723	69.3	700	64.5	N.R.	N.R.
Households that added and lost members..	393	5.0	305	4.5	88	8.1	N.R.	N.R.

Source: Special tabulations of the 1967 Survey of Economic Opportunity basic data records. N.R. - Not Relevant.

TABLE 2.--MATCHED HOUSEHOLDS THAT CHANGED IN COMPOSITION BY TYPE OF CHANGE, COLOR OF HEAD, AND POVERTY STATUS IN 1967

(Numbers in Thousands)

Matched Households by Type of Change in Composition	Total Households		White Head Households		Nonwhite Head Households		Below Poverty Line	
	Number	Percent	Number	Percent	Number	Percent	Number	Percent
Households That Changed	7,898	100.0	6,813	100.0	1,085	100.0	1,046	100.0
Had Members Join	2,475	31.3	2,090	30.7	385	35.5	214	20.5
1 joined	2,228	28.2	1,908	28.0	320	29.5	182	17.4
2 joined	204	2.6	156	2.3	48	4.4	26	2.5
3+ joined	43	--	26	--	17	1.6	6	--
Had Members Leave	5,423	68.7	4,723	69.3	700	64.5	832	79.5
1 left	4,485	56.8	3,980	58.4	505	46.5	692	66.2
2 left	662	8.4	554	8.1	108	10.0	88	8.4
3+ left	276	3.5	189	2.8	87	8.0	52	5.0

Source: Special tabulations of 1967 Survey of Economic Opportunity basic data records

-- Less than 1.0 percent

TABLE 2a.--MATCHED HOUSEHOLDS THAT CHANGED IN COMPOSITION BY NUMBER OF PERSONS WHO JOINED OR LEFT

(Numbers in Thousands)

Households That Changed In Composition By Number of Persons Who Joined or Left	Total Households	Households That Had Members Join	Households That Had Members Leave	Households That Had Members Join and Leave
Total Households	7,898	2,475	5,423	393
1 person changed	6,713	2,228	4,485	297
2 persons changed	866	204	662	49
3+ persons changed	319	43	276	48
PERCENT DISTRIBUTION				
Total Households	100.0	100.0	100.0	100.0
1 person changed	85.0	90.0	82.7	75.6
2 persons changed	11.0	8.2	12.2	12.5
3+ persons changed	4.0	1.7	5.1	12.0

Source: Special tabulations of 1967 Survey of Economic Opportunity basic data records.

TABLE 3.--MATCHED HOUSEHOLDS THAT CHANGED IN COMPOSITION CLASSIFIED BY COLOR OF HEAD, TYPE OF CHANGE, AND REASON OF CHANGE

(Numbers in Thousands)

Matched Households by Type of Change in Composition and Reason for Change	Total Households		White Head Households		Nonwhite Head Households	
	Number	Percent	Number	Percent	Number	Percent
Matched Households with a Change......	7,898	N.R.	6,813	N.R.	1,085	N.R.
Reason children joined, - total.....	1,907	100.0	1,645	100.0	261	100.0
Born since February 1966..........	1,483	77.8	1,324	80.5	159	60.9
Parents moved in for job..........	39	2.0	29	1.8	11	4.2
Parents were divorced............	74	3.9	52	3.2	22	8.4
Death in the family..............	14	--	11	--	3	1.1
Returned from an institution......	13	--	12	--	1	--
Other............................	283	14.8	218	13.3	66	25.3
Reason Adults joined, - total.......	1,229	100.0	914	100.0	287	100.0
Moved in to take a job............	157	12.8	120	12.8	38	13.2
To get married...................	213	17.3	170	18.1	42	14.6
Returning from the Armed Forces...	37	3.0	33	3.5	4	1.4
Divorced.........................	110	9.0	91	9.7	19	6.6
Death in the family..............	86	7.0	60	6.4	26	9.1
Return from an institution........	22	17.9	16	1.7	6	2.1
Other............................	603	49.1	450	47.8	153	53.3

Source: Special tabulations of the 1967 Survey of Economic Opportunity basic data records

N.R. - Not Relevant

-- Less than 1.0 percent

TABLE 3a.--MATCHED HOUSEHOLDS THAT CHANGED IN COMPOSITION BY COLOR OF HEAD AND REASON FOR CHANGE

(Numbers in Thousands)

Reason Person Left Household	Total Number of Leavers		White Heads		Nonwhite Heads	
	Number	Percent	Number	Percent	Number	Percent
Reason person left - total	5,259	100.0	4,670	100.0	589	100.0
To get married	762	14.5	652	14.0	110	18.7
Entered Armed Forces	1,686	32.1	1,533	32.8	153	26.0
Divorced	842	16.0	764	16.4	78	13.2
Died	423	8.0	335	7.2	89	15.1
Entered institution	1,328	25.3	1,196	25.6	132	22.4
Other	219	4.1	190	4.1	28	4.8

Source: Special tabulations of the 1967 Survey of Economic Opportunity basic data records.

TABLE 4.--MATCHED HOUSEHOLDS BY NUMBER OF PERSONS WHO JOINED OR LEFT THE HOUSEHOLD BY TYPE OF CHANGE AND INCOME OF PERSONS

(Numbers in Thousands)

Income of Joiners and Leavers	Matched Households with Income Contributing Members	Households that Added Members		Households that Lost Members	
		Number	Percent	Number	Percent
Number of households containing joiners or leavers.............	6,635	819	N.R.	5,816	N.R.
Total number of joiners and leavers........................	8,164	878	100.0	7,286	100.0
Level of income for leavers and joiners					
None...........................	846	379	43.2	467	6.0
$ 1.00 to $1,499...............	6,289	188	21.4	6,101	83.7
$1,500 to $2,999...............	538	118	13.4	420	5.8
$3,000 to $4,999...............	205	101	11.5	104	1.4
$5,000 or more.................	179	49	5.6	130	1.8
N.A...........................	104	39	4.4	65	.9

Source: Special tabulations of the 1967 Survey of Economic Opportunity basic records.

N.R. - Not Relevant.

TABLE 5.--PERCENT OF FAMILY INCOME CONTRIBUTED BY PERSONS WHO ENTERED OR LEFT THE MATCHED HOUSEHOLDS

Percent of Family Income Contributed	Total	Family Income		
		Under $2,000	$2,000 to $4,999	$5,000 and over
Total households........	6,852	898	1,603	4,351
0 percent.........	4,887	694	1,161	3,032
1 - 9 percent.........	309	17	52	240
10 - 19 percent.........	279	12	48	219
20 - 29 percent.........	288	16	68	204
30 - 39 percent.........	194	5	41	148
40 - 49 percent.........	152	1	46	105
50 - 59 percent.........	120	10	27	83
60 - 69 percent.........	94	18	28	48
70 - 79 percent.........	87	1	45	41
80 - 89 percent.........	77	5	12	60
90 - 99 percent.........	365	119	75	171
VERTICAL PERCENTS				
Total households........	100.0	100.0	100.0	100.0
0 percent.........	71.3	77.2	72.4	69.6
1 - 9 percent.........	4.5	1.8	3.2	5.5
10 - 19 percent.........	4.0	1.3	2.9	5.0
20 - 29 percent.........	4.2	1.7	4.2	4.6
30 - 39 percent.........	2.8	.5	2.5	3.4
40 - 49 percent.........	2.2	.1	2.8	2.4
50 - 59 percent.........	1.7	1.1	1.6	1.9
60 - 69 percent.........	1.8	2.0	1.7	1.1
70 - 79 percent.........	1.2	.1	2.8	.9
80 - 89 percent.........	1.1	.5	.7	1.3
90 - 99 percent.........	5.3	13.2	4.6	3.9

Source: Special tabulations of the 1967 Survey of Economic Opportunity basic records

SOME DETERMINANTS OF THE VARIATION IN EARNINGS FOR COLLEGE MEN

Ritchie H. Reed and Herman P. Miller,
U.S. Bureau of the Census

This study assesses the relative importance of several factors on the earnings of men with college degrees. More specifically, it examines the relationship between earnings of college men and eight other factors. They were; age, their college's rank as measured by the index of freshmen aptitude, field of specialization, color (white-nonwhite), father's occupation, father's education, current region of residence, and type of residence at high school graduation (metropolitan-nonmetropolitan). Age was included as a control variable and proved to be important, whereas, current region of residence was also intended as a control but was not shown to be important. Additional variables relating to the college attended and the student's background were eliminated either on an a priori basis or by trying them in regression equations and finding that they added little to the model's explanatory powers. The study shows how much of the variance in earnings can be explained with these variables, as well as the average earnings levels for individuals with various combinations of these characteristics.

The data were investigated by multiple regression techniques using a quantitative dependent variable (average weekly earnings) and sets of dummy values for the independent variables. Both additive and interaction models were used.

Source of the Data and Method of Analysis

The data were derived from a special supplement to the Current Population Survey of the Bureau of the Census conducted in March and April of 1967. The supplement sample consisted of males who had completed four or more years of college, and included information on each person's college major, number and level of degrees, his background, and the names of all colleges where he received a degree. Using an Office of Education code for colleges, it was possible to obtain several characteristics of the colleges attended simply by knowing their names. The information gathered in the supplement on college attendance was matched to the Current Population Survey record giving the general demographic and economic characteristics of each individual in the sample. The total sample size for men with college degrees who worked full-time was 2,559. The included 1,759 with bachelor degrees as their highest degree, with the remainder having a higher degree. Out of those with degrees, there were 115 nonwhite persons, and 2,444 whites.

The dependent variable used in this study is average weekly earnings during 1966 for full-time workers. Earnings, as defined here, included money received from wages or salaries, or from operation of a farm, business or professional practice. It does not include interest, dividends, rents or sources of income other than earnings. Since earnings are not distributed according to the normal curve, the classical regression assumptions have not been met. This factor does not bias the regression coefficients, but it does make the standard errors larger than they would be otherwise. As a result, the relationships derived will appear to be less significant than they really are.[1]

The most important independent variables were age, the college's rank as measured by the index of freshman aptitude, field of specialization, and color. Four categories were used for age, six for college rank, 11 for field of specialization, and two for color. Variables were also added for whether the father was a white-collar worker, whether the father had completed 4 years of college or more, whether the person lived in the South or not, and whether the person lived in a metropolitan area when he graduated from high school. These four characteristics proved to be relatively unimportant as explanatory variables. For each characteristic, one category was excluded to make the solution of the regression equation determinant.

The index of freshmen aptitude, or what will be referred to as college rank, is of particular interest. In the simplest terms, this index measures the average aptitude, verbal and mathematical, of entering freshmen at a given college. The data were gathered in connection with the "Project Talent" survey of high school students conducted by the University of Pittsburgh. Four cohorts of these high school students were followed up one year after graduating from high school, and the name of their college was recorded along with their measured "aptitude". The aptitude score itself is a combination of three aptitude scores--reading comprehension, abstract reasoning, and mathematics. Each school with at least 10 students in the "Project Talent" survey was given the average score of those in the survey entering that institution. These scores were then standardized to a distribution with a mean of 50 and standard deviation of 10. The measure can be interpreted as a rough indicator of the rank of a college in terms of the "ability" of the student body at a given point in time. If it is assumed that, on the average, the better students select the better colleges, this index can also be interpreted as a rough measure of the quality of a school.

The remaining independent variables are largely self-explanatory. A white-nonwhite dichotomy was used for color. The field of specialization was gathered in response to the question, "In what field did you receive this degree?"

In order to study interaction effects, the sample was divided into three levels of degrees, I, II, and III. Level I is composed of men with bachelor's degrees as the highest degree, as well as a few who stated they had a degree but did not give the level. Level II is composed of those with master's degrees, or a first degree in law, theology, or denistry. The Level III sample contained only 100 observations and contained those with the Ph.D. degree, medical and other doctors, and second degrees in law and theology. This sample was too small for our regression model, and only results for Levels I and II are presented here.

The Results

Initially the dependent variable (weekly earnings) was regressed on each characteristic separately. The results are given in table 1 for the two levels of degrees. In both cases, age, college rank, and field of specialization are the most important explanatory variables. This is shown by the relatively high values of the coefficients of determination, which measure the percent of variation in earnings that is explained by a given independent variable. Color is also relatively important. Age is the most important variable for Level I whereas college rank and color are the most important explanatory variables for Level II. On the basis of these regressions, the multivariate model was designed to include age, college rank, and field of specialization, with the remaining variables added in order of importance by a step-wise regression procedure. Father's occupation, region of residence, father's education, and type of high school residence added very little to the explanatory power of the models. In all cases, the step-wise regression added them in the order listed in table 1.

Table 1.--MEASURES OF CORRELATION AND SIGNIFICANCE FOR AVERAGE WEEKLY EARNINGS OF MEN WITH COLLEGE DEGREES AND SELECTED VARIABLES, BY DEGREE GROUP

Item	Level I		Level II	
	Coefficient of determination R^2	F	Coefficient of determination R^2	F
Age	.079	37.53*	.039	6.99*
College rank	.042	12.69*	.048	5.85**
Field of specialization	.040	7.25*	.072	5.38**
Color	.008	14.25*	.012	8.24*
Age, rank, specialization, and race combined	.134	12.73*	.180	6.76*
Father's occupation	.004	7.29*	.007	4.67**
Current region of residence	.003	5.81**	.003	2.08
Father's education	.002	2.69	.001	.40
High school residence	.002	3.94	.001	.82
All variables combined	.135	10.83*	.184	5.85*

* Significant at .01 level.
** Significant at .05 level.

College Rank

The regression results for college rank, are given in table 2. The data labeled "Gross Effect" are the results of regressing the dependent variable on college rank alone. The (b) coefficients measure the average increment to earnings due to being in a given aptitude group rather than in the lowest group, which is the excluded variable. The mean earnings levels for the excluded group is measured by the intercept. For instance, people who received bachelor's degrees from schools in group 6 had average weekly earnings of $105 more than those who attended the poorest schools, whose average earnings level was $169. The mean weekly earnings column shows the mean weekly earnings for those in the different groups. The data labeled "Net Effect" measure the effects of college rank, after the effects of age, field of specialization, and color have been accounted for. The net mean incomes are calculated under the assumption that age, college rank, field of specialization, and color are additive variables, or in other words, that the effect of each is independent of the value of the other variables.[2]

The results show that there are differences in earnings due to college rank for both degree levels. Going to a better school is more of an advantage for those with higher degrees. It is also apparent that taking account of other factors than college rank significantly affects the apparent role of the college rank variable. In general, the additional variables tend to decrease the increments in earnings associated with college rank, and to increase the intercept.

The data suggest that, on the average, going to the best schools rather than the poorest adds about $4,200 per year for the holder of a bachelor degree (Level I). This difference is greater for holders of Level II degrees--approximately $5,900. One possible explanation for the difference is that the quality of education is more important at the higher degree levels where performance on the job is more directly related to technical skills acquired at the University.

Although account has been taken of interaction between college rank and the level of degree, the regressions for each level assume additivity and this may not be a realistic representation of the data. The effect of the college's rank may be different for whites and nonwhites, or for education majors and engineering majors. Testing for these differences would have required a much larger sample.

The increments to earnings due to college rank are greater for the higher rank colleges than the lower ones, in the Level I sample. However, the college rank measure should be interpreted as having only ordinal properties and therefore nothing can be concluded about the marginal returns to attending better colleges. That is, the regression coefficients do not apply to equal intervals on the rank scale, because distance between rankings has no meaning in an absolute sense.[3]

Table 2.--THE EFFECTS OF COLLEGE RANK ON AVERAGE WEEKLY EARNINGS OF MEN WITH COLLEGE DEGREES, BY LEVEL OF DEGREE

College rank as measured by index of freshmen aptitude group	Gross effect[1]			Net effect[2]		
	b	σ_b	Mean weekly earnings in dollars	b	σ_b	Mean weekly earnings in dollars
Level I						
0		Intercept	169	Adjusted	Intercept	173
1	-5	23.5	164	13	23.2	186
2	15	23.1	184	25	23.0	198
3	39	22.1	208	35	22.0	208
4	54	22.4	223	43	22.3	216
5	84	23.8	253	65	23.6	238
6	105	24.8	274	84	24.7	257
Level II						
0		Intercept	167	Adjusted	Intercept	159
1	-7	38.3	160	21	38.0	180
2	58	39.2	225	85	37.8	244
3	63	37.2	230	75	35.6	234
4	78	37.0	245	79	35.4	238
5	75	39.0	242	58	37.2	217
6	115	39.4	282	118	37.6	277

[1]Regression of earnings on college rank.
[2]Regression of earnings on college rank with color, age, and field of specialization as control variables.

Field of Specialization

Table 3 shows the differences in average weekly earnings accounted for by the field of specialization. For Level I degrees--engineering, the physical sciences, and business and commerce offer the greatest monetary rewards. Those majoring in religion did the poorest. The pattern changes somewhat for those in the Level II sample. Health fields and law offer the greatest returns to persons in this sample with business and commerce and education improving relative to other majors. Those in the health fields are primarily dentists, as there are no MD's in the Level II sample.

Persons majoring in technical fields for a bachelor's degree appear to have an income advantage that is lost for holders of higher degrees. In the Level I sample for example, men who majored in engineering and the physical sciences received the highest returns, but for those with their highest degree in the Level II sample, business and commerce overtakes both. This may be somewhat misleading, for many people with a master's degree in business have bachelor's degrees in technical fields. In a sense, this earnings differential may be due to a difference in occupations. An individual with a second degree in engineering is very likely to be a research worker whereas an individual with a graduate degree in business is more likely to be in the higher paid levels of management. Thus, we have not entirely separated the effects of college major and current occupation.[4]

Table 3.--THE EFFECTS OF FIELD OF SPECIALIZATION ON AVERAGE WEEKLY EARNINGS OF MEN WITH COLLEGE DEGREES, BY DEGREE GROUP

Field of specialization	Gross effect[1]			Net effect[2]		
	b	σ_b	Mean weekly earnings in dollars	b	σ_b	Mean weekly earnings in dollars
Level I						
Education.................		Intercept	164	Adjusted	Intercept	182
Biological sciences.......	-7	18.4	157	-22	17.6	160
Business and commerce.....	52	11.8	216	38	11.5	220
Engineering...............	87	12.3	251	58	12.3	240
Health....................	31	23.7	195	15	22.7	197
Humanities................	32	17.2	196	16	16.6	198
Physical sciences.........	70	15.1	234	43	14.8	225
Religion..................	-21	47.5	143	-31	45.6	151
Social sciences...........	37	13.8	201	18	13.5	200
Other.....................	60	21.7	224	27	21.2	209
Level II						
Education.................				Adjusted	Intercept	190
Biological sciences.......				5	39.5	195
Business and commerce.....				74	21.8	264
Engineering...............				58	22.5	248
Health....................				107	25.8	297
Humanities................				-7	29.3	183
Law.......................				100	16.9	290
Physical sciences.........				20	24.6	210
Religion..................				-14	27.1	176
Social sciences...........				-22	24.7	168
Other.....................				-5	49.7	185

[1]Regression of earnings on field of specialization.
[2]Regression of earnings on field of specialization with color, age, and college rank as control variables.

Color

The regressions show that nonwhites have significantly lower earning levels than whites (table 4). According to the "Gross Effects" of color, there is a greater difference in earnings due to color at the higher degree level. However, this difference disappears when color is used in a multivariate model. For both levels, the "Gross Effect" attributes some factors to color that should be attributed to college rank and field of specialization. Color of the student and the rank of his institution are related as well as color and certain fields of specialization. This can be seen in table 5. Thus when these variables are added, the earnings differential due to color is decreased. But, the net effect still represents over $2,400 per year difference in earnings, even after taking account of college rank, field of specialization, age, and level of degree. The nonwhite category includes Japanese and Chinese who usually have significantly higher incomes than Negroes. Therefore the earnings differences between whites and Negroes would be greater than those shown here for whites and nonwhites.

Table 6 illustrates how adding color as a variable affects the regression coefficients for college rank. The first column gives the regression coefficients for college rank, controlling for field of specialization and age. The addition of color to the regression changed the regression coefficients to those in the third column. For both degree levels, the introduction of color increases the intercept value and decreases each regression coefficient by a relatively uniform amount.

The size of the change in the regression coefficients indicates that there is substantial correlation between color and college rank. Without color, the regression attributed earnings effects to college rank which are more properly attributed to color. The pattern of the changes in the college rank regression coefficients indicate that being nonwhite and in a low ranking college are likely concurrences. What originally appeared as earning increments associated with attending schools above the lowest rank is actually due to being white as opposed to nonwhite.

It is important to see how different variables may reinforce the effects of color for Negroes. After accounting for age, college rank, and field of specialization, nonwhites have significantly lower earnings than whites. As can be seen from table 5, Negroes are more likely to choose a field of specialization that is less rewarding financially, (education) and less likely to be engineers, the most rewarding for bachelor degree holders. Add to this the fact that such a large proportion of Negroes attend the lowest rank colleges. And, the Negro age structure is such that Negroes are more likely to be young if they have a college degree. By each factor, the Negro has a disadvantage and none of the factors offset others, but rather they reinforce each other. Even though $49 of the difference in white and nonwhite weekly earnings is due to color, the actual difference in mean weekly earnings was $67. The difference between $67 and $49 is due to other factors than color. Being nonwhite, and the attributes, choices and opportunities that are associated with being nonwhite, caused nonwhite earnings to be 51 percent of white earnings for men with bachelor's degrees.

Table 4.--THE EFFECTS OF COLOR ON THE AVERAGE WEEKLY EARNINGS OF COLLEGE MEN, BY LEVEL OF DEGREE

Color	Gross effect[1]			Net effect[2]		
	b	σ_b	Mean weekly earnings in dollars	b	σ_b	Mean weekly earnings in dollars
Level I						
White		Intercept	214	Adjusted	Intercept	214
Nonwhite	-67	17.7	147	-49	17.5	165
Level II						
White		Intercept	232	Adjusted	Intercept	231
Nonwhite	-83	29.1	149	-48	27.9	183

[1]Regression of earnings on color.
[2]Regression of earnings on color with college rank, age, and field of specialization as control variables.

Table 5.--DISTRIBUTION OF SAMPLE AMONG COLLEGE RANKS AND FIELD OF SPECIALIZATION FOR MEN WHOSE HIGHEST DEGREE IS A BACHELOR DEGREE, TOTAL AND NEGRO

College rank	Total	Negro	Field of specialization	Total	Negro
Total, all ranks	100.0	100.0	Total, all fields	100.0	100.0
0	2.8	24.0	Education	13.5	34.0
1	11.7	36.0	Biological sciences	5.0	4.0
2	14.6	6.0	Business and commerce	25.2	18.0
3	30.4	8.0	Engineering	20.6	4.0
4	22.6	16.0	Health	2.6	8.0
5	10.6	6.0	Humanities	6.0	0.0
6	7.3	4.0	Physical sciences	9.0	14.0
			Religion	0.6	0.0
			Social sciences	12.5	14.0
			Other	3.2	0.0
			Not reported	1.8	4.0

Table 6.--CHANGES IN SIZE OF REGRESSION COEFFICIENTS FOR INDEX OF FRESHMEN APTITUDE DUE TO THE ADDITION OF "COLOR" AS A CONTROL VARIABLE

Degree group and rank of college as measured by the index of freshmen aptitude	Regression coefficient without color[1]	Increment in mean earnings over preceding aptitude group in dollars	Regression coefficient with color[2]	Increment in mean earnings over preceding aptitude group in dollars
Level I				
Intercept	46	--	63	--
1	26	26	13	13
2	41	15	25	12
3	50	9	35	10
4	59	9	43	8
5	81	22	65	22
6	100	19	84	19
Level II				
Intercept	55	--	65	--
1	26	26	21	21
2	92	66	85	64
3	80	-12	75	-10
4	86	6	79	4
5	62	-24	58	-21
6	122	60	118	60

[1]Regression of earnings on college rank, with age and field of specialization as control variables.

[2]Regression of earnings on college rank, with color, age, and field of specialization as control variables.

Background variables

Earlier studies have demonstrated that background factors are important in determining a person's educational attainment.[5] Whether or not the parents have a college education has been shown to be important in determining if a child will get a college education or degree. The regression analysis presented here suggests that among college graduates, the educational attainment and occupation of the family head does not have a demonstrable effect on earnings. Of the two variables, it was found that father's occupation has a stronger influence than the father's education even though neither were important. Whether or not a man lived in a metropolitan area when he graduated from high school proved to be of minimal importance.

CONCLUSION

Using all of the variables we have considered, including the background variables, about 13.5 percent of the variation in earnings for college men with bachelor's degrees and 18.4 percent for men with Level II degrees can be explained. Using college rank, age, field of specialization, and race we can explain 13.4 percent of earnings variation for Level I and 18.0 percent for Level II. There are, of course, many factors which affect earnings that have not been taken into account in this analysis. More of the variation in earnings could have been explained if it were possible to measure factors such as physical and mental health, personality, ambition, and intelligence.

Earnings appear to be positively related to college rank, as measured by the index of freshman aptitude. In addition, some choices of major in college are clearly superior to others in terms of earnings potential. However, the same majors are not the most promising at all degree levels. Furthermore, nonwhites have an earnings disadvantage even after accounting for the rank of their college, their age, and college major. For Negroes, all of the factors we have considered, work cumulatively to lower their earnings as compared to whites. The family background variables studied seem to have almost no effect on earnings.

ACKNOWLEDGMENTS

The authors wish to thank George Heller and Steven H. Schwartz of the Census Bureau for their invaluable guidance and help in computer processing of the data.

FOOTNOTES

[1]Another factor may counteract or overwhelm this underestimation of significance. Since the Current Population Survey is taken from a statified sample, the measures of variance derived for statistics from this sample will tend to be smaller than these that would be derived from a random sample.

FOOTNOTES--Continued

[2]The adjusted intercept for college rank is calculated as the sum of the multivariate intercept and the products of the regression coefficients, for age, field of specialization, and race, and the proportion of the sample falling in each category of these variables. Net mean earnings for college rank are then derived as the sum of the adjusted intercept and the appropriate regression coefficient for college rank. Thus the net mean earnings data indicate the effects of college rank after the effects of age, field of specialization, and race have been accounted for. Net mean incomes for age, field of specialization, and race are calculated similarly.

[3]Indeed, the regression coefficients might be said to measure the cardinal distances between aptitude groups. If one chooses to define college quality in terms of the relative earning power of graduates from the different rank colleges, we are measuring college quality rather than testing a hypothesis that relative college quality is related to earnings.

[4]Using either field of specialization or occupation alone makes the results biased in the sense that earnings effects would be allocated to one factor when both share in their determination. If we had included occupation as a variable, the regression coefficients for the several fields of specialization would have been smaller and probably less statistically significant. However, for the questions being asked here, occupation is not useful as a variable. Since one's college major influences his occupation and not vice versa, some of the effects that would be attributed to occupation would actually be due to college major. Thus, if occupation were included in the regressions, the effects of one's choice of a college major would be underestimated.

An additional factor to consider is the assumption of additivity in the regressions for field of specialization. It is shown below that there is a significant interaction effect between age and field of specialization. Typically, an engineer or a physicist has an initial earnings advantage over a lawyer or a doctor, but the same relationship does not hold for persons in middle or late career. Therefore, to come extent the earnings differentials due to age and field of specialization are not being separated from one another by this analysis.

MEDIAN INCOMES IN SELECTED OCCUPATIONS FOR PERSONS WITH 5 OR MORE YEARS OF COLLEGE, BY AGE: 1960

Selected occupation	Age (years)			
	25 to 34	35 to 44	45 to 54	55 to 64
Lawyers and judges...	7,272	12,157	14,636	13,635
Electrical engineers.	9,010	11,529	11,142	10,145
Physicians & surgeons	4,866	19,663	21,048	17,664
Physicists...........	8,808	11,676	(NA)	(NA)

Source: U.S. Bureau of the Census, U.S. Census of Population: 1960, Subject Reports, Occupation by Earnings and Education, Final Report PC(2)-7B, U.S. Government Printing Office, Washington, D.C., 1963.

[5]Brazer, Harvey E., and Martin David, "Social and Economic Determinants of the Demand for Higher Education," in U.S. Department of Health, Education, and Welfare, Office of Education, Economics of Higher Education (Washington, D.C., U.S. Government Printing Office, 1962), pp. 21-42.

XIII

CONTRIBUTED PAPERS III

STATISTICAL METHODS: NEW APPLICATIONS AND DEVELOPMENTS

Chairman, MARGARET E. MARTIN, U. S. Office of Statistical Policy

RANKING COLLEGE BASKETBALL TEAMS

M. L. Goodman, Bell Telephone Laboratories, Incorporated, Murray Hill, New Jersey

I. Introduction

The problem of ranking college basketball teams belongs to the class of problems associated with analyzing the results of a paired comparisons experiment; that is, an experiment in which a set of objects are compared two at a time, and the better one in each pair is identified. The reader should bear in mind that the results presented in this paper are applicable to the general paired comparisons experiment although the paper is written in the terminology of sports, using such words as "team" and "game" instead of "object" and "comparison".

There is a considerable literature to draw upon for analyzing paired comparisons experiments, but there are two major reasons why most published methods are inappropriate for ranking college basketball teams. First, most of the methods are directed towards the balanced tournament in which every team plays every other team the same number of times. College basketball is very unbalanced. Secondly, most of the methods are oriented towards selecting the best team in the tournament or in making tests of hypothesis about the equality of different sets of teams. In the present case, we are concerned with obtaining an ordered ranking of all the teams.

There is also a wide range in the amount of information which can be incorporated into a ranking algorithm. One may restrict oneself to utilizing only the knowledge of the better (winning) team as is the case with the methods of this paper. As an extension, one might include information on the measure of difference between the two teams. This is usually accomplished by including the score of the game as part of the data, and such methods tend to be oriented toward regression or analysis of variance techniques.

Finally, one might include in the data various additional measures obtained for the individual members of each team. Of course, the more measures one introduces, the greater is the danger that an unexpected interaction between the measures will be introduced which may render the rankings invalid. For example, in many sports, hockey in particular, an "all-star" game is played in which the first-place team plays against an "all-star" team composed of the best players from the remaining teams in the league. Almost any algorithm which used measures based on the players' performances would rank the all-star team ahead of the first-place team - yet the first-place team seldom loses such games. This is usually attributed to such causes as teamwork and spirit, which are impossible to include in an objective ranking scheme, but certainly do exist.

In light of these comments, it seems reasonable to define three desirable properties that any objective ranking scheme should contain.

First, a scheme should restrict itself to <u>objective</u> data. For ranking sports, this means that only the results of games should be used and no modification should be included because a key player may have had an upset stomach. The amount of objective data to be used may vary, but one should keep in mind the aforementioned danger of including too much.

Secondly, the scheme should be <u>impartial</u>. This means that if a team has a certain rank then any other team with the identical record (with respect to the algorithm) should have the exact same rank.

Finally, the scheme should be <u>directionally invariant</u>. This property, which was originally proposed by W. A. Larsen [4], means that if a ranking of all the teams has been computed and then two teams play one additional game, the rank of the winning team shall not decrease, nor shall the losing team's rank increase.

II. Currently Used Methods

At present, college basketball teams are ranked by the two major news services, the Associated Press (AP) and the United Press International (UPI). They derive the ranking by polling voters: the voters in the UPI poll are a consistent panel of 35 coaches, while the panel for the AP poll consists of a varying group of 35 to 45 sportswriters.

Prior to 1968, both polls asked each voter to name what he considered the top ten teams. Every first place vote received 10 points, a second place vote earned 9 points, and so on. The polls totaled the votes received by each team and then released a list of the top ten vote getters, the so-called top ten teams. They also released a list of all other teams receiving any votes.

In 1968, the AP initiated an arbitrary elaboration on the simple scheme: voters were asked to list the top 15 teams. Points were assigned on a 20-18-16-14-12-10-9-8-7-6-5-4-3-2-1 basis, and a list of the top 20 teams were released. (The UPI, while continuing to vote on only ten teams, also began listing the top 20.)

The unusual point assignment used by the AP does not lack precedent; for example, in balloting for the most valuable player in baseball, each voter ranks ten men, and a weighting system of 14-9-8-7-...-1 is used.

The methods currently in use suffer from the serious defect that the identities of the voters will influence the outcome of the poll. For example, a coach in the UPI poll can virtually assure his own team being ranked in the top 20 by ranking it fourth or fifth himself. A fourth place ranking gives a team 7 points and a team with 7 points will usually be ranked between 16th and 19th in the UPI poll. Such a situation seemed to exist in the 1968-69 UPI basketball poll when a well-known Western school received six or seven points almost every week even though their record was mediocre, and, moreover, they received no points at all in the AP poll. This meant that no writer thought this school belonged in the top 15, but some coach thought they were fourth. Of course, this defect is partly due to a system which ranks 20 teams when only 10 are voted on, and it would appear no matter what weights were used. One way to combat it might be to prevent any coach from voting on his own team and then multiplying the total votes for his team by N/(N-1).

A more serious aspect of the same defect is that no coach or writer can watch every good team play. Consequently, their vote will partly reflect the few schools that they have actually seen play. For the rest, they will consider a team's record; specifically, whom they have beaten and to whom they have lost.

Here an interesting proposition arises: since much of the ranking is already done on the basis of whom a team has played, it seems reasonable to create an impartial, formal mathematical method for doing so. The remainder of this paper will discuss the problems in creating such a method and will describe a method which seems to yield a reasonable ranking of the college basketball teams.

III. Forerunners

Our first attempt to develop a ranking algorithm commenced with an investigation of a method proposed by Wei [5] and published by Kendall [3]. This method is based on the hypothesis that the winning team in a round robin tournament should not necessarily be the one with the most victories; rather teams should get more credit for beating good teams than poor ones. To accomplish this end the method uses as ranks the eigenvector corresponding to the largest eigenvalue of the won-lost matrix, W. This matrix is defined by its elements: w_{ij} = the number of times team i has beaten team j. Since the eigenvector has the property that premultiplying it by the matrix is the same as multiplying it by a constant, the value assigned in the ranking to any particular team may be seen to be the normalized sum of the values of the teams it has defeated. Thus a team does get more credit for defeating a good team. It was the eigenvector corresponding to the largest eigenvalue that was used because the method as originally developed used an iterative procedure to determine the rankings, and when this method is applied to the original won-lost matrix, it is the largest eigenvector which is obtained. Additional statistical implications and interpretations of using this specific eigenvector need to be studied further.

This ranking method was applied in Kendall's original paper to only the results of a balanced, round-robin tournament. We, of course, want to apply it to a very unbalanced case. To do so, data was gathered for the 1968-69 college basketball season on 191 teams. The details of the data gathering are given in the Appendix.

One important consideration was how to include games played by schools in the group of 191 against schools not in the group. The magnitude of the problem may be seen in Table 1 which contains the distributions of total games played and non-group games played by the schools in the study. It may be seen that while the average team played 25 games, only three were against non-group opponents, and only four schools played more than half their games against such opposition. Several modifications were tried to account for these games; none of them seemed to have a serious effect on the rankings; and the results of such games have been omitted from the rankings included in this paper.

A sample of the results of directly applying Wei's method to the 1968-69 college basketball data is given in the K-score column of Table 2. These may be seen to differ substantially from the AP and UPI rankings for the same week. Moreover, several schools with mediocre records are included in the K-score's top 20, and all of the teams belong to major conferences; in fact, 17 of the top 20 belong to just four conferences, the Big Ten, Big Eight, Pacific Eight, and Western Athletic. Wei's method may thus be seen to inflate the ranks of conference schools at the expense of independents; the supreme example of this for the week of March 2 is that Stanford of the Pacific Eight Conference with a won-lost record of 8-17 was ranked 25th while LaSalle, one of the top Eastern independents, was ranked 59th, even though their record was 23-1.

It was felt that perhaps one of the reasons that Wei's method performed poorly in the unbalanced case was that while it was ranking teams on the basis of whom they have beaten, it was completely ignoring all information about to whom they had lost. This is not a serious drawback in the balanced case since all teams play the same number of games against the same opponents and losses are considered in that they are games which are not won. Such is not the situation, however, in the unbalanced case; hence an alternative method was developed which utilizes information on both wins and losses.

IV. Proposed Method

This alternative method, which is similar in nature to Wei's, attempts to account simultaneously both for whom a team has lost to as well as for whom a team has beaten. To do this it assumes that every team has some underlying value. For a given team this value is the normalized sum of the values of the teams it has beaten minus a correction for the teams to whom it has lost. For each loss, this correction is computed to be $V_{max} - V_L$, where V_{max} is the value of the top ranked team and V_L is the value of the team to whom you lost.

The method may be most easily understood by referring to formulas (4.1) to (4.4). Thus, the value of the jth team, V_j, is determined as

$$V_j = U_j/K \, , \qquad (4.1)$$

where

$$K = \frac{2}{N} \sum_{i=1}^{N} U_i \, , \qquad (4.2)$$

and

$$U_j = T_j - T_{min} \, , \qquad (4.3)$$

where

$$T_j = \sum_{\substack{W=\text{all} \\ \text{teams} \\ \text{beaten}}} V_W - \sum_{\substack{L=\text{all} \\ \text{teams} \\ \text{lost to}}} (V_{max}-V_L) \qquad (4.4)$$

$$= \sum_{\substack{p=\text{all} \\ \text{teams} \\ \text{played}}} V_p - N_L V_{max}$$

where N_L = number of games lost.

An iterative procedure is used to determine the rankings derived by this method. Thus, if at some point in the iterative process, we have a vector of ranking values, $\underline{V}$, then the value for the jth team in the next iteration is computed as follows: First obtain the T_j value which is equal to the sum of the values of the teams defeated minus the correction for teams lost to as in equation (4.4). Next, the bottom ranked team is constrained to have a value of 0; this is accomplished by the subtraction in equation (4.3). Finally, in order to achieve convergence, the sum of the values for all teams is constrained to equal N/2. The value N/2 was chosen because the sum of the teams' won-lost percentages will also approximately equal N/2; it was felt that this approximate equality would enable comparisons to be made between a team's percentage and its ranking value. In any case, the constant, K, of equation (4.2) is the divisor necessary to yield the correct sum, and equation (4.1) merely indicates the division.

Under this method it may be seen that a team gets no credit for beating the poorest team in the country, but loses nothing for losing to the best. In addition, when two good teams play, the one who loses is penalized very little in relation to all other teams (since $V_{max}-V_L$ will be small), while the one who wins gains quite a bit. The converse is true for a game between two poor teams.

Finally, the method quickly iterates to a vector of stable values. Using the college basketball data and using the won-loss percentages as the first vector of values in the iterative process, the author has found the five digit accuracy is obtained within 20 iterations.

The results of applying this method to the March 2 data are given in the G-score column of Table 2. It may be seen that these ranks do bear some resemblance to those produced by the news services, and the author would, of course, argue in favor of the G-score rankings because, as will be discussed in Section V, these rankings must closely satisfy the criteria set forth at the beginning of the paper. In addition, the final rankings, including all post-season tournament games, for all 191 teams are given in Table 3.

V. Summary

This paper has discussed three possible methods for ranking teams in badly unbalanced tournaments, and the methods have been applied to college basketball. It seems appropriate to examine how well the three methods conform to the three criteria for ranking algorithms set forth in the beginning of the paper. These were (i) objectivity, (ii) impartiality, and (iii) directional invariance.

The news service polls, as discussed earlier, are neither objective nor impartial. Due to the psychology of the voters they are likely to have directional invariance, but this property cannot be proved.

Wei's method is both objective and impartial but is seriously deficient in directional invariances as several test cases have shown.

The alternative method is also both objective and impartial. Tests have shown that it is also directionally invariant except in one rare case where a team defeats the poorest team in the country having already defeated that team earlier in the season. In such a case, the winning team adds nothing to its T value in equation (4.4) but the losing team has its T value decreased by $V_{max} - V_L$. In equation (4.3), T_{min} is now smaller than before; hence the U and V values of all teams are increased. Now, since the worst team is lower ranked, relative to the other teams, than it was before, the teams which have earlier beaten this poorest team have their T values increased less than the average. Hence, the winning team in the game just played shows a less than average increase due to the poorest team having lost again, and gets nothing for its latest win. It thus can fall in the rankings. As mentioned above, however, such cases are extremely rare. This seems a small price to pay for a method which seems to have so many desirable properties, including the ability to rank more than one undefeated team.

Finally, it seems worthwhile to ask how this method behaves in the completely balanced case for which Wei's method was originally developed. Tests on sample data have shown that in the completely balanced case the ranking values derived by the proposed method are merely a linear transformation of the won-loss percentages. This result leads this author to conclude that these percentages yield the most reasonable ranks in the balanced case while the proposed algorithm should be applied in very unbalanced situations.

APPENDIX

The college teams that have been included in this ranking are those whose complete schedules were available to the author prior to the start of the 1968-69 college basketball season. The schedules of 190 of the teams were found in the Basketball Yearbook [2]. The schedule for Long Island University was given in the New York Times [1] and this school was included in the rankings because it played most of its games against schools already in the group.

Most of the results of the games were called by the author from the daily sports pages of the New York Times. The author would like to thank Mr. James Blinn and Mrs. Ione Breyer who regularly made available sports sections from the Des Moines Register and the Chicago Tribune, respectively. These papers often contained scores of games not reported in the Times.

Finally, the results of all games for all teams in the study are not included in the ranking. This is because the author was unable to obtain the results of many games played in the Rocky Mountains and Far West, due to the somewhat parochial orientation of his sources which were located in the East and Midwest. This incompleteness of results has no effect on the ranking algorithm described in this paper; it does mean, however, that the ranks reported in this paper are not the absolute, final season rankings for the 1968-69 college basketball season. To remedy this, the author would greatly appreciate any readers who can supply an accurate record of schools whose records as given in this paper are incorrect or incomplete.

REFERENCES

[1] "Basketball Schedules for Colleges in the Metropolitan Area and other Eastern Teams," New York Times, Sports Section, November 24, 1968.

[2] Hollander, Zander, Ed., Basketball Yearbook 1969, Popular Library, Inc., New York, 1968.

[3] Kendall, M. G., "Further Contributions to the Theory of Paired Comparisons," Biometrics, 11 (1955), 43-62.

[4] Larsen, W. A., personal communication.

[5] Wei, T. H., The Algebraic Foundations of Ranking Theory, unpublished thesis, Cambridge University, England, 1952.

Total Games Played

# Games	# Teams
18	1
19	3
20	1
21	4
22	6
23	14
24	40
25	36
26	40
27	18
28	13
29	10
30	3
31	1
32	1

Median = 25
Mean = 25.06

Games Played Against Non-Group Teams

# Games	# Teams
0	30
1	25
2	27
3	32
4	16
5	14
6	12
7	6
8	9
9	7
10	3
11	5
12	1
13	2
14	1
15	1

Median = 3
Mean = 3.72

TABLE 1

DISTRIBUTIONS OF GAMES PLAYED

	K-Score	W	L	Score	AP	W	L	Points*	UPI	W	L	Points**		G-Score	W	L	Score
1.	UCLA	24	0	2.798	UCLA	24	0	898	UCLA	24	0	350	1.	UCLA	24	0	1.056
2.	Purdue	18	4	2.159	LaSalle	23	1	724	Santa Clara	24	1	254	2.	North Carolina	22	3	.967
3.	Kansas	20	5	1.909	Santa Clara	24	1	650	North Carolina	22	3	244	3.	Davidson	25	2	.940
4.	Colorado	19	6	1.865	North Carolina	22	3	606	Davidson	25	2	204	4.	St. John's	22	4	.931
5.	Ohio State	15	7	1.731	Davidson	25	2	573	LaSalle	23	1	193	5.	Villanova	21	4	.905
6.	North Carolina	22	3	1.719	Purdue	18	4	465	Purdue	19	4	173	6.	LaSalle	23	1	.901
7.	Illinois	17	5	1.691	Kentucky	20	4	306	Kentucky	21	4	141	7.	Purdue	18	4	.867
8.	Northwestern	13	9	1.645	St. John's	22	4	335	St. John's	22	4	92	8.	Santa Clara	24	1	.857
9.	Washington State	17	8	1.544	Duquesne	19	3	292	Duquesne	19	3	44	9.	Kentucky	20	4	.855
10.	Kentucky	20	4	1.511	Villanova	21	4	203	Villanova	20	4	44	10.	Drake	21	4	.849
11.	Brigham Young	16	11	1.511	Drake	21	4	159	Drake	21	4		11.	Duquesne	19	3	.847
12.	Michigan	13	9	1.367	New Mexico State	23	2	154	New Mexico State	23	2		12.	New Mexico State	23	3	.839
13.	Wyoming	20	7	1.349	South Carolina	19	5	122	Wyoming	20	7		13.	Notre Dame	20	5	.836
14.	USC	14	11	1.342	Marquette	21	4	119	Notre Dame	20	5		14.	South Carolina	19	5	.823
15.	Missouri	14	9	1.335	Louisville	21	4	102	Colorado	19	6		15.	Illinois	17	5	.819
16.	Drake	21	4	1.306	Boston College	20	3	85	South Carolina	19	5		16.	Boston College	20	3	.818
17.	Michigan State	11	10	1.288	Notre Dame	20	5	61	Marquette	12	4		17.	Kansas	20	5	.789
18.	Utah	14	13	1.284	Colorado	19	6	46	Kansas	20	5		18.	Louisville	18	4	.787
19.	Iowa State	13	12	1.282	Kansas	20	5	38	Boston College	20	3		19.	Ohio State	15	7	.767
20.	Kansas State	12	12	1.260	Illinois	17	5	27	Princeton	19	6		20.	Dayton	20	5	.767

TABLE 2

RANKINGS INCLUDING GAMES OF MARCH 2

* 45 writers vote for 15 teams. Points are given on a 20-18-16-14-12-10-9-8-7-6-5-4-3-2-1 basis.

** 35 coaches vote for 10 teams. Points are given on a 10-9-8-7-6-5-4-3-2-1 basis.

TABLE 3

FINAL RANKINGS FOR 1968-9 COLLEGE BASKETBALL SEASON

RANK		W	L	T	PCT.	SCORE
1	U C L A	29.	1.	0.	0.967	1.2377848
2	NORTH CAROLINA	27.	5.	0.	0.844	1.0385638
3	DAVIDSON	27.	3.	0.	0.900	1.0085019
4	PURDUE	23.	5.	0.	0.821	1.0065071
5	ST. JOHN'S	23.	6.	0.	0.793	0.9632869
6	DRAKE	26.	5.	0.	0.839	0.9338918
7	BOSTON COLLEGE	24.	4.	0.	0.857	0.9292973
8	LA SALLE	23.	1.	0.	0.958	0.9288516
9	KENTUCKY	23.	5.	0.	0.821	0.9250950
10	WEBER STATE	26.	3.	0.	0.897	0.9127226
11	DUQUESNE	21.	5.	0.	0.808	0.9087441
12	VILLANOVA	21.	5.	0.	0.808	0.8932999
13	SANTA CLARA	27.	2.	0.	0.931	0.8862888
14	ILLINOIS	19.	5.	0.	0.792	0.8756470
15	OHIO STATE	17.	7.	0.	0.708	0.8312137
16	TEMPLE	22.	8.	0.	0.733	0.8277592
17	SOUTH CAROLINA	21.	7.	0.	0.750	0.8218129
18	NEW MEXICO STATE	24.	5.	0.	0.828	0.8210571
19	MARQUETTE	24.	5.	0.	0.828	0.7870573
20	COLUMBIA	20.	4.	0.	0.833	0.7835201
21	RUTGERS	21.	4.	0.	0.840	0.7787215
22	LOUISVILLE	21.	6.	0.	0.778	0.7780006
23	NOTRE DAME	20.	7.	0.	0.741	0.7769433
24	TENNESSEE	21.	7.	0.	0.750	0.7752372
25	PRINCETON	19.	7.	0.	0.731	0.7725692
26	WASHINGTON STATE	18.	8.	0.	0.692	0.7599324
27	COLORADO	21.	7.	0.	0.750	0.7508259
28	KANSAS	20.	7.	0.	0.741	0.7351411
29	NORTHWESTERN	14.	10.	0.	0.583	0.7306671
30	WAKE FOREST	18.	9.	0.	0.667	0.7136241
31	DAYTON	20.	7.	0.	0.741	0.7019833
32	FLORIDA	18.	9.	0.	0.667	0.6949419
33	TULSA	20.	8.	0.	0.714	0.6837789
34	MURRAY STATE	22.	6.	0.	0.786	0.6762792
35	ST. BONAVENTURE	17.	7.	0.	0.708	0.6664945
36	ARMY	18.	10.	0.	0.643	0.6660492
37	MASSACHUSETTS	17.	7.	0.	0.708	0.6630674
38	ST. PETER'S	21.	7.	0.	0.750	0.6614033
39	COLORADO STATE	18.	7.	0.	0.720	0.6485279
40	WYOMING	20.	9.	0.	0.690	0.6452762
41	TEXAS A & M	18.	9.	0.	0.667	0.6439517
42	FLORIDA STATE	18.	8.	0.	0.692	0.6385160
43	BAYLOR	18.	6.	0.	0.750	0.6375750
44	SOUTHERN CALIF.	15.	12.	0.	0.556	0.6372300
45	MICHIGAN	13.	11.	0.	0.542	0.6361748
46	FORDHAM	17.	9.	0.	0.654	0.6333660
47	NORTH CAROLINA ST.	15.	10.	0.	0.600	0.6292224
48	HOLY CROSS	17.	8.	0.	0.680	0.6284317
49	OHIO UNIVERSITY	17.	9.	0.	0.654	0.6278469
50	MONTANA STATE	17.	8.	0.	0.680	0.6216159
51	ARIZONA	17.	10.	0.	0.630	0.6213655
52	VANDERBILT	15.	11.	0.	0.577	0.6210283
53	SEATTLE	19.	9.	0.	0.679	0.6134436
54	ST. JOSEPH'S	17.	11.	0.	0.607	0.6118016
55	WESTERN KENTUCKY	16.	9.	0.	0.640	0.6077598
56	CINCINNATI	17.	9.	0.	0.654	0.6061935

57	WEST TEXAS STATE	16.	8.	0.	0.667	0.5992673
58	DUKE	15.	14.	0.	0.517	0.5985894
59	AUBURN	15.	10.	0.	0.600	0.5972895
60	MOREHEAD STATE	20.	9.	0.	0.690	0.5844282
61	HOUSTON	16.	10.	0.	0.615	0.5800981
62	PACIFIC	17.	9.	0.	0.654	0.5789539
63	MINNESOTA	12.	12.	0.	0.500	0.5782612
64	PROVIDENCE	14.	10.	0.	0.583	0.5776248
65	PENNSYLVANIA STATE	13.	9.	0.	0.591	0.5751085
66	WISCONSIN	11.	13.	0.	0.458	0.5750757
67	LONG ISLAND UNIV	17.	6.	0.	0.739	0.5716541
68	MICHIGAN STATE	11.	12.	0.	0.478	0.5711356
69	CREIGHTON	13.	13.	0.	0.500	0.5693349
70	NEW MEXICO	17.	9.	0.	0.654	0.5631830
71	EASTERN KENTUCKY	13.	8.	0.	0.619	0.5607526
72	MIAMI (OHIO)	15.	12.	0.	0.556	0.5588999
73	SOUTHERN ILLINOIS	17.	7.	0.	0.708	0.5586438
74	DETROIT	16.	10.	0.	0.615	0.5561265
75	OREGON	13.	13.	0.	0.500	0.5556530
76	PENNSYLVANIA	15.	10.	0.	0.600	0.5521900
77	IOWA	12.	12.	0.	0.500	0.5511747
78	UC - SANTA BARBARA	17.	9.	0.	0.654	0.5501344
79	BRIGHAM YOUNG	17.	12.	0.	0.586	0.5490079
80	NEW YORK UNIV.	12.	9.	0.	0.571	0.5482714
81	ST. FRANCIS (PA)	14.	8.	0.	0.636	0.5458268
82	KANSAS STATE	14.	12.	0.	0.538	0.5423048
83	TEXAS EL PASO	15.	9.	0.	0.625	0.5355758
84	MIAMI (FLA)	14.	9.	0.	0.609	0.5292407
85	WASHINGTON	13.	13.	0.	0.500	0.5254429
86	BOSTON UNIVERSITY	14.	10.	0.	0.583	0.5232376
87	JACKSONVILLE	16.	7.	0.	0.696	0.5216934
88	SAN JOSE STATE	16.	8.	0.	0.667	0.5165821
89	MANHATTAN	13.	9.	0.	0.591	0.5136680
90	EAST CAROLINA	17.	11.	0.	0.607	0.5096606
91	IOWA STATE	14.	13.	0.	0.519	0.5026182
92	GETTYSBURG	14.	9.	0.	0.609	0.5005942
93	OKLAHOMA CITY	20.	9.	0.	0.690	0.4993614
94	GEORGIA	13.	12.	0.	0.520	0.4984426
95	NORTH TEXAS STATE	15.	10.	0.	0.600	0.4982091
96	KENT STATE	14.	10.	0.	0.583	0.4966845
97	CALIFORNIA	12.	13.	0.	0.480	0.4934951
98	GONZAGA	8.	11.	0.	0.421	0.4900799
99	EAST TENNESSEE ST.	16.	11.	0.	0.593	0.4900774
100	INDIANA	9.	15.	0.	0.375	0.4849725
101	UTAH	14.	13.	0.	0.519	0.4848205
102	LOUISIANA STATE U	13.	13.	0.	0.500	0.4842134
103	MISSOURI	14.	11.	0.	0.560	0.4830031
104	DE PAUL	14.	11.	0.	0.560	0.4773446
105	OREGON STATE	12.	14.	0.	0.462	0.4699722
106	VIRGINIA TECH	14.	12.	0.	0.538	0.4639436
107	TOLEDO	13.	11.	0.	0.542	0.4636083
108	MAINE	11.	11.	0.	0.500	0.4630665
109	SOUTHERN METHODIST	12.	12.	0.	0.500	0.4529476
110	HARDIN-SIMMONS	12.	13.	0.	0.480	0.4496715
111	BUTLER	12.	14.	0.	0.462	0.4491024
112	GEORGE WASHINGTON	14.	11.	0.	0.560	0.4489108
113	NORTHERN ILLINOIS	13.	12.	0.	0.520	0.4377969
114	ARIZONA STATE	11.	15.	0.	0.423	0.4314986
115	OKLAHOMA STATE	13.	13.	0.	0.500	0.4285234
116	VIRGINIA	10.	15.	0.	0.400	0.4271492
117	BRADLEY	14.	12.	0.	0.538	0.4239027

118	WEST VIRGINIA	13.	14.	0.	0.481	0.4235923
119	IDAHO	7.	12.	0.	0.368	0.4218751
120	NEBRASKA	12.	14.	0.	0.462	0.4183628
121	NIAGARA	11.	13.	0.	0.458	0.4153466
122	TENNESSEE TECH	12.	11.	0.	0.522	0.4148331
123	GEORGETOWN	12.	12.	0.	0.500	0.4121817
124	C C N Y	3.	16.	0.	0.158	0.4120749
125	PEPPERDINE	14.	11.	0.	0.560	0.4104573
126	HOFSTRA	12.	13.	0.	0.480	0.4097295
127	RIDER	11.	14.	0.	0.440	0.4062102
128	BUCKNELL	13.	11.	0.	0.542	0.4030226
129	GEORGIA TECH	12.	13.	0.	0.480	0.4025187
130	IONA	11.	11.	0.	0.500	0.3981723
131	MIDDLE TENN. STATE	13.	13.	0.	0.500	0.3976493
132	TULANE	11.	14.	0.	0.440	0.3966306
133	MISSISSIPPI	10.	14.	0.	0.417	0.3857803
134	WICHITA STATE	12.	15.	0.	0.444	0.3856080
135	STANFORD	9.	17.	0.	0.346	0.3814375
136	VERMONT	12.	11.	0.	0.522	0.3803636
137	TEXAS CHRISTIAN	13.	12.	0.	0.520	0.3764197
138	SETON HALL	9.	16.	0.	0.360	0.3763733
139	MARSHALL	9.	15.	0.	0.375	0.3763352
140	FAIRFIELD	10.	16.	0.	0.385	0.3756861
141	WESTERN MICHIGAN	11.	13.	0.	0.458	0.3712312
142	DELAWARE	8.	10.	0.	0.444	0.3701057
143	NAVY	7.	14.	0.	0.333	0.3683732
144	MISSISSIPPI STATE	8.	17.	0.	0.320	0.3637952
145	COLGATE	11.	14.	0.	0.440	0.3629565
146	CORNELL	12.	13.	0.	0.480	0.3620943
147	FAIRLEIGH DICKINSON	10.	14.	0.	0.417	0.3594347
148	LOYOLA (ILL)	9.	14.	0.	0.391	0.3570189
149	RICE	10.	14.	0.	0.417	0.3499307
150	IDAHO STATE	6.	14.	0.	0.300	0.3465961
151	MARYLAND	8.	18.	0.	0.308	0.3393085
152	RHODE ISLAND	10.	15.	0.	0.400	0.3369106
153	MONTANA	5.	16.	0.	0.238	0.3361670
154	AUSTIN PEAY	9.	14.	0.	0.391	0.3350544
155	RICHMOND	13.	14.	0.	0.481	0.3346985
156	XAVIER	10.	16.	0.	0.385	0.3244307
157	TEXAS TECH	11.	14.	0.	0.440	0.3241227
158	THE CITADEL	13.	12.	0.	0.520	0.3175245
159	SYRACUSE	9.	16.	0.	0.360	0.3027223
160	NEW HAMPSHIRE	9.	15.	0.	0.375	0.3005763
161	AIR FORCE	11.	13.	0.	0.458	0.2986615
162	CANISIUS	7.	16.	0.	0.304	0.2837577
163	TEXAS	9.	15.	0.	0.375	0.2831118
164	UTAH STATE	10.	17.	0.	0.370	0.2824042
165	YALE	9.	16.	0.	0.360	0.2794774
166	BOWLING GREEN	9.	15.	0.	0.375	0.2771720
167	ST. FRANCIS (NY)	7.	16.	0.	0.304	0.2765337
168	ARKANSAS	10.	14.	0.	0.417	0.2687935
169	CLEMSON	7.	19.	0.	0.269	0.2616466
170	LEHIGH	6.	17.	0.	0.261	0.2545359
171	ST. MARY'S	6.	19.	0.	0.240	0.2436783
172	DARTMOUTH	10.	15.	0.	0.400	0.2378562
173	LAFAYETTE	9.	17.	0.	0.346	0.2279705
174	CENTENARY	8.	19.	0.	0.296	0.2250639
175	ST. LOUIS	6.	20.	0.	0.231	0.2189118
176	OKLAHOMA	7.	19.	0.	0.269	0.2172865
177	SAN FRANCISCO	8.	18.	0.	0.308	0.1945456
178	HARVARD	7.	18.	0.	0.280	0.1899031

179	FURMAN	9.	17.	0.	0.346	0.1800938
180	PORTLAND	2.	20.	0.	0.091	0.1765330
181	MEMPHIS STATE	6.	19.	0.	0.240	0.1738683
182	PITTSBURGH	4.	20.	0.	0.167	0.1704313
183	VIRGINIA MIL INST	5.	18.	0.	0.217	0.1657655
184	LOYOLA (LA)	5.	19.	0.	0.208	0.1488945
185	CONNECTICUT	5.	19.	0.	0.208	0.1480441
186	ALABAMA	4.	20.	0.	0.167	0.1399336
187	LOYOLA (CAL)	5.	19.	0.	0.208	0.1285840
188	AMERICAN UNIV.	4.	19.	0.	0.174	0.1192185
189	WILLIAM & MARY	6.	20.	0.	0.231	0.0871988
190	DENVER	2.	24.	0.	0.077	0.0034623
191	BROWN	3.	23.	0.	0.115	0.

QUASI-EXPERIMENTAL DESIGN: FURTHER EXPLORATIONS

Robert F. Boruch and Leroy Wolins
American Council on Education   Iowa State University

Campbell (1965) has loosely defined quasi-experimental design as the application of the experimental mode of analysis to behavioral science situations which do not meet complete requirements of classical experimental control or design. He has focused primarily on inference problems associated with nonrandom assignment of experimental units to treatments. Other examinations of the quasi-experimental paradigm in the behavioral sciences have taken the form of analytic concern with relaxation of constraints in classical design models, and augmentation of common analyses to assay or compensate for deviations from classical assumptions. Such research has resulted in some interesting information about the relations between experimental classificatory models and the factor analytic models more commonly used by social scientist. For example, Gollob (1968) has presented a method of examing data which is based on a two-way (mixed) analysis of variance model in order to account for correlated error. The usual analytic procedure is augmented by a principal components representation of matrix of interaction parameters. Bock and Bargmann (1966) have conducted research which is based on systematic variation of the general factor analysis model. Structural characteristics of some models are related to the usual experimental designs. Maximum likelihood rather than least squares techniques are used to estimate parameters in this case, however.

In this paper, we shall limit consideration to deriving a general relation between a mixed classificatory model with fewer constraints and a restricted factor analytic model. Demonstration and evaluation of the model usage is based on simulated data.

Consider now the situation characterized by the following attributes: (a) A mixed analysis of variance model is with random main effect hypothesized, but alternative models (with fewer classical restrictions) are plausible; (b) Primary interest lies in the random main effects and interactions, rather than in the fixed effects.

These conditions are rather common in psychological and educational research. Students are frequently the random effect, and treatments or blocking attributes can be interpreted as the fixed effects. Insofar as fixed effect scales are arbitrary in the social sciences, there is often little justification for examining absolute values of scale scores. Rather, the interaction of student and score are most informative to the investigator.

The classical model and assumptions are given in equation (1) of the Appendix: a three-way ANOV representation with a random main effect. Sources of variance involving the random effects and random interactions can be examined conveniently by obtaining a covariance matrix in which the summation is taken over the random effect (subscript i). Assuming that the sample is large and using the assumptions indicated, the expected covariances can be condensed to expression (5). Multipliers of the variance terms are Kronecker deltas.

If the fixed effect scales are arbitrary, one can equate the averages to some constant value so that fixed main effects and the fixed interaction are unaffected by the procedure. One can show that adjustment of the observed scores, Y_{ijk} to y_{ijk}, does not alter expected values of the covariances (Stanley, 1961).

Suppose we now relax the assumption of unit weighting of each random effect in this model. Instead, the factors can be weighted differentially, depending on the specific jk combination (7). It is not unreasonable to conjecture that the weights attached to effects are a function of a specific treatment-block combination, for example. If the classical assumptions are true, the A_{jk}, B_{jk}, C_{jk} can be interpreted as the variance components associated with random effects. When parameters are dependent on jk, the squared estimates provide something <u>like</u> variance components within a jk combination.

Given this last equation, the substitutions indicated in equation (8) are made in order to conform to common factor analysis model notation. The expectation of the dispersion matrix under this model is given in equation (11), and matrix definitions are given in Appendix II for the case of j=1,2,3, k=I,II,III.

This is of the form of the usual factor analytic model considered by Lawley(1940) and others. It is restricted in the sense that certain factor structure and factor correlation matrix elements are constrained by the investigator to be zero, and the others are free and must be estimated.

Note that the dimensions and attributes of the factor structure and factor correlation matrix are a function of the original experimental design. In order to assure a meaningful solution, one must attend to the uniqueness of the hypothesized factor structure. Anderson and Rubin (1956) and Koopmans and Riersol (1950) provide sufficient conditions for indentification (up to rotation) of the solution. For parsimonious data description and for easy interpretation of results, the correlations among the factors (X_i, X_{ij}, X_{ik}) are usually assumed to be zero. In this orthogonal case, the <u>phi</u> matrix is an identity and the factor structure given is unique.

Estimation of Parameters

The maximum likelihood estimates of free parameters in the restricted factor analysis model can be obtained by using an extension of Lawley's (1940) method. Joreskog and Gruvaeus (1967) provide detailed description of the procedure, and an excellent computer program for its implementation. One can impose on the hypothesized factor structure and on the factor correlation matrix <u>a priori</u> constraints that certain parameters are exactly equal to zero. This restricted maximum likelihood factor analysis (RMLFA) allows the investigator to make a large sample, Chi-square test for goodness of fit of the model to the data.

The procedure recommended by Joreskog (1967) for evaluation of goodness of fit is a sequential one. After successively altering the factor structure hypothesized, one accepts the solution having the best fit at the probability level specified. Since the tests are conditonal, one knows only that the probability of the accepted structure being different from the true structure is less than or equal to the specified significance level. Joreskog further suggest random division of the sample into halves, using one half to generate a final hypothesis of interest in a sequential manner, and the other half to test the hypothesis. His recommendations refer largely to testing hypotheses about the number of factors in the unrestricted (i.e., no fixed elements) factor analytic model. Lacking information to the contrary, we shall assume that much the same procedure is appropriate in making sequential decisions about modification of the restricted model.

Simulation of Data

In order to assess the utility of the model and of the sequential procedure described earlier, a pilot Monte Carlo study was initiated. The four basic factor structures considered are provided in Table 1; attention is restricted to orthogonal structures only.

Random floating point numbers, distributed NID (0,1) were generated and used as the independent variates (i.e., factor scores) in the factor analytic model. Sample size is restricted to be 200, and number generation was accomplished using a computer program (Control Data 3600) developed by Wolfe (1968). Two samples, each consisting of 200 units, comprise the validation sample (for a test of the final hypothesis). Linear functions of the random variables were computed on the basis of population factor loadings indicated in the matrices of Table 1 and the model (8) given in the Appendix. Variance-covariables were derived from the resulting sample observations.

The generated independent random variables conform closely to normal curve frequencies. Chi-square tests of hypotheses that the variance-covariance matrix for the independent varialbes is a sample from a population whose covariance matix is a diagonal results in acceptance of the null hypothesis. The hypothesis of independence among the observations (i.e., the linear combinations) is rejected for each generated matrix.

The form of the population factor structure for synthesis of samples conforms to requirements for uniqueness of the solution (Joreskog and Gruvaeus, 1967). Magnitudes of the factor loadings, and generated correlations are similar to those commonly obtained in analyses of psychological data. Structures 1 and 4 were chosen because they are suggestive of the experimental design situation characterized by homogeneity of error variance and additivity of effects. Structures 2 and 3 are minor variations on this model.

Results

Models which are successively hypothesized to fit the data are represented by factor patterns given in Table 2. Each pattern matrix corresponds to a hypothesized factor structure, with 1's representing free parameters which must be estimated, and 0's representing the elements which are restricted to be zero. For the first factor structure, four models were hypothesized, the last one (D) being the true model. Five models were each hypothesized to fit the data generated on the basis fo structures 2,3, and 4. The last model (E) in each case is the true one. The models increase in complexity to simulate the investigatior's objective to obtain parsimony in description.

Several attributes of any intermediate (false) solution warrant attention. Changes in successive hypothesized factor structures can be made conditional on such characteristics, in order to achieve a better fit of the model to data. Boruch and Wolins (in press) and Joreskog (1967) supply detailed information, including examples, on some of the following criteria.

Boundary problems occur when estimates of parameters fall at or outside the region of allowable values. Such improper solutions may be acceptable (Joreskog, 1967). The solution is not a maximum likelihood solution, since partial derivatives at the solution defined by parameter estimates are not all zero.

The size and significance of the Chi-square statistic, associated with maximum likelihood factor analysis, is an appropriate index of the goodness of fit in the confirmatory sense. The magnitude of the statistic, when used in successive exploratory tests of hypotheses, is a convenient index for examining the goodness of fit. In making comparisons of successive solutions in which degrees of freedom differ the computed Chi-square, divided by degrees of freedom, may be useful. The division allows examination of mean square residuals adjusted for degrees of freedom.

If multiple independent samples are available and hypothesized factor structures are planned in advance, then the ratios of independent Chi-squares, divided by degrees of freedom, can be examined.

Other devices for summarizing the results of particular analysis are commonly used: examination of residual correlations, of consistency of an estimated factor correlation matrix (in the oblique solution). For restricted solutions considered here, near zero estimates of parameters are also of interest.

Consider now the summary data provided in Table 3. Boundary problems are designated in the third column of the chart. In three instances (all similar factor structures), the limiting value of specific error variances was met. The hypothesized model was rejected on this basis. If the boundary condition is ignored, the magnitude of the Chi-square would lead to acceptance of the model.

The Chi-square tests for <u>false</u> models lead appropriately to rejection of the hypothesis in all cases except IV (Model D), while the confirmatory sample test (ICV) leads to a marginal rejection. The ratio of the Chi-square for models C and D to their respective degrees of freedom is not enlightening. The adjustment for degrees of freedom fails to show which model might be more acceptable in an umambiguous way.

The rejection of true models in independent tests appears to occur more frequently than one would expect (1 CV-Model D, 2 CV-Model E). A Type I error also occurs for the solution 2V-Model E. The rejections are marginal, but suggest that the probability of a Type I error may be larger than advertized. A tendency toward rejection of true solutions is part of the anecdotal rather than the systematic information in unrestricted factor analysis studies based on limited samples. These results imply that a similar problem affects tests on restricted models.

Examination of the estimated factor Loading for each solution reveals values near zero for each false hypothetical model. The values occur frequently for situations in which parameters are hypothesized but the true factor structure does not contain the elements.

The ratio of Chi-square statistic appears to be useful for interpretation of results. In general, the ratio decreases as the hypothesized models are successively changed to conform more closely to the true structure of the data. In the case of a comparison of Models B and C conditional on the factor structurel, for example, this is the case. Without rather well defined, systematic methods of successive testing, however, this index of fit is not likely to be very useful.

The ratios of successive Chi-square divided by their respective degrees of freedom are not appropriate for evaluation with respect to significance levels, since the successive Chi-square values are not independent. The ratios, however, are indicative of the magnitude of improvement in fitting two successive models. Of course, the more drastic changes in hypothesized factor structure are associated with the larger ratios.

This cursory examination is informative in only a suggestive way. Although the notion of exploratory and confirmatory factor analytic techniques appears to be appropriate for an experimental design-like situation as described earlier, the actual results achieved in this small situation are a bit dubious. There is some evidence to suggest that the probability of accepting a true solution is somewhat larger than the tabulated likelihoods, at least for the structures considered.

If one has no plan for sustematic alteration of hypothesized models, it is unlikely that successive tests to the true model will be as straightforward as the procedures demonstrated here. With some such system and the use of the Chi-square and Chi-square/degree of freedom ratio, some reasonable results can be achieved.

References

Anderson, T.W. and Rubin, H. Statistical inference in factor analysis. Proceedings of the Third Berkeley Sumposium on Mathematical Statistics and Probability, 1956, 111-150.

Bock, R.D. and Bargmann, R. Analysis of covariance structures. Psychometrika, 1966, 4, 507-534.

Boruch, R.F. and Wolins, L. A procedure for estimatingthe amount of method, trait and error variance attributable to a measure. Educational and Psychological Measurements. (in press)

Campbell, D.T. Quasiexperimental design. Proceedings of the American Statistical Association. Social Statistics Section, 1965, 157-160.

Gollob, H.F. A statistical model which combines features of factor analytic and analysis of variance techniques. Psychometrika, 1968, 33, 73-115.

Joreskog, K.G. Some contributions to maximum likelihood factor analysis. Psychometrika, 1967b, 32, 443-482.

Joreskog, K.G. and Gruvaeus, G. RMLFA, a computer program for restricted maximum likelihood factor analysis. Research Bulletin RB-67-21, Educational Testing Service, Princeton, New Jersey, 1967a.

Koopmans, T.C. and Riersol, O. Identification of structural characteristics. Annals of Mathematical Statistics. 1950, 21, 165-181.

Lawley, D. N. The estimation of factor loadings by the method of maximum likelihood. Proceedings of the Royal Society of Edinburgh, Section A, 1940, 60, 64-82.

Stanley, J. C. Analysis of unreplicated three way classifications, with applications to rater bias and trait independence. Psychometrika, 1961, 26, 205-219.

Wolfe, R.G. Random floating point mumbers in Normal Distribution. Unpublished manuscript, University of Wisconsin Computing Center, Madison 1965.

TABLE 1

Population Factor Structures

(1)

700
700
700
700
700
700
700
700
700

(2)

70	00	00	50	00	00
70	00	00	00	40	00
70	00	00	00	00	65
00	44	00	50	00	00
00	44	00	00	40	00
00	44	00	00	00	65
00	00	55	50	00	00
00	00	55	00	40	00
00	00	55	00	00	65

(3)

77	00	00	58	00	00
66	00	00	00	32	00
71	00	00	00	00	62
00	45	00	50	00	00
00	49	00	00	40	00
00	44	00	00	00	65
00	00	58	44	00	00
00	00	61	00	44	00
00	00	47	00	00	70

(4)

70	00	00	50	00	00
70	00	00	00	5C	00
70	00	00	00	00	50
00	50	00	50	00	00
00	50	00	00	50	00
00	50	00	00	00	50
00	00	70	50	00	00
00	00	70	00	50	00
00	00	70	00	00	50

TABLE 2

Factor Patterns for Sequential Testing of Models

For Factor Structure **(1)**:

A

1	0	0
1	0	0
1	0	0
0	1	0
0	1	0
0	1	0
0	0	1
0	0	1
0	0	1

B

1	0	0	1
1	0	0	0
1	0	0	0
0	1	0	1
0	1	0	0
0	1	0	0
0	0	1	1
0	0	1	0
0	0	1	0

C

1	1	0	0
1	1	0	0
1	1	0	0
1	0	1	0
1	0	1	0
1	0	1	0
1	0	0	1
1	0	0	1
1	0	0	1

D

1
1
1
1
1
1
1
1
1

For Factor Structures (2), (3), and (4):

A

1	0	0
1	0	0
1	0	0
0	1	0
0	1	0
0	1	0
0	0	1
0	0	1
0	0	1

B

1	0	0	1
1	0	0	0
1	0	0	0
0	1	0	1
0	1	0	0
0	1	0	0
0	0	1	1
0	0	1	0
0	0	1	0

C

1	0	0	1	0
1	0	0	0	1
1	0	0	0	0
0	1	0	1	0
0	1	0	0	1
0	1	0	0	0
0	0	1	1	0
0	0	1	0	1
0	0	1	0	0

D

1	1	0	0	1	0	0
1	1	0	0	0	1	0
1	1	0	0	0	0	1
1	0	1	0	1	0	0
1	0	1	0	0	1	0
1	0	1	0	0	0	1
1	0	0	1	1	0	0
1	0	0	1	0	1	0
1	0	0	1	0	0	1

E

1	0	0	1	0	0
1	0	0	0	1	0
1	0	0	0	0	1
0	1	0	1	0	0
0	1	0	0	1	0
0	1	0	0	0	1
0	0	1	1	0	0
0	0	1	0	1	0
0	0	1	0	0	1

Appendix I

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + \ell_{ijk} \quad (1)$$

where

$i = 1,2,3, \ldots, N$ $\qquad \ell_{ijk} \sim NID(0,\sigma^2_\varepsilon)$

$j = 1,2,3, \ldots, b$ $\qquad \alpha_i \sim NID(0,\sigma^2_\alpha)$

$k = 1,2,3, \ldots, c$ $\qquad [(\alpha\beta)_{ij} \sim NID(0,\sigma^2_{\alpha\beta})]$

$[(\alpha\gamma)_{ik} \sim NID(0,\sigma^2_{\alpha\gamma})]$.

$$\underset{i}{Cov}(Y_{ijk},Y_{ij'k'}) = E\{(Y_{ijk} - E(Y_{.jk}))(Y_{ij'k'} - E(Y_{.j'k'}))\} \quad . \quad (2)$$

where

$E(\alpha_i) = 0$ $\qquad E(\underset{i}{\Sigma}(\alpha\beta)_{ij}) = 0$ $\quad$ (3)

$E(\beta_j) = \beta_j$ $\qquad E(\underset{i}{\Sigma}(\alpha\gamma)_{ik} = 0$.

$E(\gamma_k) = \gamma_k$

$E((\beta\gamma)_{jk}) = (\beta\gamma)_{jk}$

$E(\underset{i}{\Sigma}\varepsilon_{ijk}) = 0$.

$$E\{(Y_{ijk}-E(Y_{.jk}))(Y_{ij'k'} - E(Y_{.j'k'}))\} \quad (4)$$

$$= E\{(\alpha_i + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + \varepsilon_{ijk})(\alpha_i + (\alpha\beta)_{ij'} + (\alpha\gamma)_{ij'} + \varepsilon_{ij'k'})\}$$

$$\underset{i}{Cov}(Y_{ijk},Y_{ij'k'}) = \sigma^2_\alpha + \delta_{jj'}\sigma^2_{\alpha\beta} + \delta_{kk'}\sigma^2_{\alpha\gamma} + \delta_{jj'}\delta_{kk'}\sigma^2_\varepsilon \quad . \quad (5)$$

$$y_{ijk} = \alpha_i + \alpha\beta_{jk} + \alpha\gamma_{jk} + \varepsilon_{ijk} \quad . \quad (6)$$

$$y_{ijk} = A_{jk}\alpha_i + B_{jk}(\alpha\beta)_{ij} + C_{jk}(\alpha\gamma)_{ik} + \varepsilon_{ijk} \quad (7)$$

A_{jk} , B_{jk} , C_{jk} are parameters where

$$\varepsilon_{ijk} \sim NID(0,\sigma^2_{\varepsilon jk})$$

$$\alpha_i \sim NID(0,1)$$

$$(\alpha\beta)_{ij}\ ,\ (\alpha\gamma)_{ik} \sim N(0,1) \quad .$$

$$y_{i(jk)} = A_{jk}X_i + B_{jk}X_{ij} + C_{jk}X_{ik} + \varepsilon_{i(jk)} \quad . \tag{8}$$

$$\rho_{jk,j'k'} = E\{(y_{ijk} - E(y_{.jk}))(y_{ij'k'} - E(y_{.j'k'}))\} \quad .$$

$$\underset{i}{Cov}(y_{ijk},y_{ij'k'}) = E\{(y_{ijk})(y_{ij'k'})\} \tag{9}$$

$$= E\{(A_{jk}\alpha_i + B_{jk}(\alpha\beta)_{ij} + C_{jk}(\alpha\gamma)_{ik} + \varepsilon_{ijk})$$

$$(A_{j'k'}\alpha_i + B_{j'k'}(\alpha\beta)_{ij'} + C_{j'k'}(\alpha\gamma)_{ik'} + \varepsilon_{ij'k'})\}$$

$$= A_{jk}A_{j'k'} + B_{jk}B_{j'k'}\rho^{BB}_{jj'} + B_{jk}C_{j'k'}\rho^{BC}_{jk'}$$

$$+ B_{j'k'}C_{jk}\rho^{CB}_{j'k'} + C_{jk}C_{j'k'}\rho^{CC}_{kk'} + \rho_{e_{jk}e_{j'k'}} \tag{10}$$

where $\rho^{BB}_{jj'}$ is, for example, the correlation between $(\alpha\beta)_{ij}$ and $(\alpha\beta)_{ij'}$.

$$\underset{i}{Cov}(y_{ijk},y_{ijk}) = \Sigma = \Lambda\Phi\Lambda' + \Psi \tag{11}$$

$$\Lambda = \begin{bmatrix} A_{II} & B_{II} & 0 & 0 & C_{II} & 0 & 0 \\ A_{21} & 0 & B_{21} & 0 & C_{21} & 0 & 0 \\ A_{31} & 0 & 0 & B_{31} & C_{31} & 0 & 0 \\ A_{1II} & B_{1II} & 0 & 0 & 0 & C_{1II} & 0 \\ A_{2II} & 0 & B_{2II} & 0 & 0 & C_{2II} & 0 \\ A_{3III} & 0 & 0 & B_{3II} & 0 & C_{3II} & 0 \\ A_{1III} & B_{1III} & 0 & 0 & 0 & 0 & C_{1III} \\ A_{2III} & 0 & B_{2III} & 0 & 0 & 0 & C_{2III} \\ A_{3III} & 0 & 0 & B_{3III} & 0 & 0 & C_{3III} \end{bmatrix}$$

ψ = Diagonal matrix with non-zero entries of $\sigma^2_{\varepsilon jk}$

NEGATIVE UNADJUSTED SUM OF SQUARES IN THE ANALYSIS OF VARIANCE

Daniel E. Schooley and Michael C. Holen, University of Oregon

The problem of a negative unadjusted sum of squares in the analysis of variance is no mere academic question to the social scientist when one suddenly appears in the analysis of his pet project. Often the experience is, at best, confusing and frustrating, leading to hours of recalculation and literature research. This paper was developed as an attempt to clarify the problem for the experimentally-minded social scientist and as an admonition to those who might be tempted to develop experiments too recklessly. A simple proof of negative unadjusted sums of squares is presented in a format typical of the statistical references commonly used by many social scientists. The authors believe the proof to be original although certainly the subject has been more elegantly treated by Anscombe (1), Nelder (2), or Thompson (3). These articles, however, are found in the Journal of the Royal Statistical Society, Biometrika, and the Annals of Mathematical Statistics. Likely not even the most optimistic among statisticians would claim that these journals are widely read, used as resource references, or even understood by most social science researchers.

The authors became aware of the problem when a graduate student was completing his thesis in social psychology. He was utilizing a four-factor, completely crossed factorial design which we will use as an example in the proof. While running his analysis, a negative sum of squares occurred in a triple interaction term. His thesis committee was skeptical (as thesis committees often tend to be) and he was summarily returned to the calculator; the negative sum of squares, however, persisted. The following proof was developed as a result of the unwillingness of that negative estimate to disappear.

The authors feel the notational system used to be consistent with those often encountered and used by social scientists [e.g., Winer (4), Myers (5)]. The terms necessary to derive an ABC interaction with at least three main factors are:

(1) $$T = \frac{\left(\sum_i^n \sum_j^a \sum_k^b \sum_o^c Y\right)^2}{n\,a\,b\,c},$$

(2) $$A = \frac{\sum_j^a \left(\sum_i^n \sum_k^b \sum_o^c Y\right)^2}{n\,b\,c},$$

(3) $$B = \frac{\sum_k^b \left(\sum_i^n \sum_j^a \sum_o^c Y\right)^2}{n\,a\,c},$$

(4) $$C = \frac{\sum_o^c \left(\sum_i^n \sum_j^a \sum_k^b Y\right)^2}{n\,a\,b},$$

(5) $$AB = \frac{\sum_j^a \sum_k^b \left(\sum_i^n \sum_o^c Y\right)^2}{n\,c},$$

(6) $$AC = \frac{\sum_j^a \sum_o^c \left(\sum_i^n \sum_k^b Y\right)^2}{n\,b},$$

(7) $$BC = \frac{\sum_k^b \sum_o^c \left(\sum_i^n \sum_j^a Y\right)^2}{n\,a},$$

(8) $$ABC = \frac{\sum_j^a \sum_k^b \sum_o^c \left(\sum_i^n Y\right)^2}{n},$$

and

(9) k, p = constants

The sums of squares for terms (2) through (8) are found in the following manner:

(10) $SS_A = A-T$,

(11) $SS_B = B-T$,

(12) $SS_C = C-T$,

(13) $SS_{AB} = AB-A-B+T$,

(14) $SS_{AC} = AC-A-C+T$,

(15) $SS_{BC} = BC-B-C+T$,

and

(16) $SS_{ABC} = ABC-AB-AC-BC+A+B+C-T$.

The inequality necessary to generate a negative sum of squares for the ABC interaction is

(17) $ABC+A+B+C < AB+AC+BC+T$

Conditions observed in the student's thesis and required for deriving the above inequality are:

(18) terms (1) through (9) $\geq$ o,

(19) a minimum of three main factors,

(20) unequal cell n's,

(21) $A \leq T+k$,

(22) C+(k+p) < BC

(23) ABC+B < AB+AC+p

Adding inequalities yields:

(24) ABC+B+A+C+(k+p) < AB+AC+BC+T+(k+p)

Adding a negative (k+p) to each side of the inequality produces the original inequality (17):

(17) ABC+A+B+C < AB+AC+BC+T.

In the student's data the double interactions achieved a magnitude outweighing the contribution of the other components. Unequal cell sizes seem of particular importance in producing the negative sum of squares because of the concomitant increase in the probability of heterogeneity of variance, a violation of a basic, and too often casually treated, assumption in the analysis of variance. The occurrence of the negative estimate is an indication that the design was inappropriate and that the data should have been analyzed differently, or better, that the experiment should have been more carefully planned and conducted.

The authors hope this paper will save future researchers hours of fruitless searching for nonexistent errors in calculation. More importantly they hope it will encourage more care in the design and analysis of experiments. While such techniques as the Doolittle method of least squares analysis, which minimize the interaction terms, tend to eliminate the possibility of negative sums of squares, there is likely no substitute for more rigorous attention to the assumptions underlying the analysis.

REFERENCES

(1) Anscombe, F. J. (1948) Contribution to the discussion on D. G. Champernowne's "Sampling theory applied to autoregressive sequences." Journal Royal Statistical Society, B, 10, 239.

(2) Nelder, J. A. (1954) The interpretation of negative components of variance. Biometrika, 41, 544-8.

(3) Thompson, W. A. Jr. (1962) The problem of negative estimates of variance components. Annals of Mathematical Statistics, 33, 273-89.

(4) Winer, B. J. (1962) Statistical principles in experimental design, New York, McGraw-Hill.

(5) Myers, J. L. (1966) Fundamentals of experimental design, Boston, Allyn and Bacon.

(6) Wald, A. (1940) A note on the analysis of variance with unequal class frequencies. Annals of Mathematical Statistics, 11, 96-100.

PREDICTION USING POOLED REGRESSIONS

Ivor Francis, Cornell University and Samprit Chatterjee, New York University

Introduction

An applicant for admission to college is usually required to take a series of tests in support of his application. In many instances the applicant's scores are then weighted and added to form one composite score, and this score will determine the success or failure of his application. The reasoning behind this procedure is that this composite score is presumed to relate to, and therefore can be used to predict, ability to succeed in college.

A psychologist might argue, however, that potential applicants form not one population, but rather they should be regarded as comprising several populations each having distinct psychological and cultural properties. This being the case it would be better, as far as uniform accuracy of prediction is concerned, to use a different set of weights for different populations rather than the same weights for everybody.

We shall suppose that it is not possible to measure directly those psychological and cultural factors which would place a candidate into one of these populations. The question is how should we estimate his ability to succeed? Several possibilities suggest themselves. Firstly we can use the same weights for everybody, as we are supposing is currently done. But if we knew the best weights for each population perhaps we could use the observed scores for an individual to first classify him into one of the populations and then use the weights that are best for that population. Alternatively, if we know the probability π_i that the new individual came from the i-th population, $i = 1, \ldots, k$, perhaps we should weight the separate population estimates with these probabilities.

The problem, then, is the following: from each of k populations we have independent sample observations on the variables $(y, x_j, j = 1, \ldots, p)$, the size of the sample in the i-th population being n_i, $i = 1, \ldots, k$. In addition we have observations on (x_j) for an individual known to have come from one of the k populations, but from which one is not known. We wish to estimate the value of y for this new individual.

First we shall present a maximum likelihood solution and then a discussion of methods of pooling the regressions of the separate populations.

The Maximum Likelihood Approach

Suppose we have k populations, with n_i observations from the i-th population, where $\Sigma n_i = n$. Let the observations on (x_j) for the new observation be denoted by $\underline{X}^*$, a $(p \times 1)$ vector. The Y-value corresponding to $\underline{X}^*$ will be denoted by Y^*. Assume y is a scalar random variable and $\underline{x}$ is a $(p \times 1)$ vector random variable with the following distribution in the i-th population

$$y \sim N(\underline{x}^t \beta_i, \sigma^2)$$

$$\underline{x} \sim N(\underline{\mu}_i, \underline{\Sigma}), \quad i = 1, \ldots, k$$

where β_i and μ_i are $(p \times 1)$ parameter vectors, and $\underline{\Sigma}$ is a $(p \times p)$ matrix, $\underline{\Sigma}$ being assumed the same in all populations.

Let $\underline{Y}_i$ denote the $(n_i \times 1)$ vector of observations on y in the i-th population, and let $\underline{X}_i$ denote the $(p \times n_i)$ matrix of observations on $\underline{x}$ in the i-th population.

In the manner of Hartley and Rao (1968) we introduce an indicator vector $\underline{\theta}$, where

$$\underline{\theta} = (\theta_1, \ldots, \theta_k) \quad \text{such that } \Sigma\, \theta_i = 1$$

where $\theta_i = 1$ if the new individual came from the i-th population and $\theta_i = 0$ otherwise.

The likelihood function for β_i, μ_i, Σ, σ^2, $\underline{\theta}$, and Y^* is

$$L(\beta_i, \mu_i, \Sigma, \sigma^2, \underline{\theta}, Y^*) = L_1 \times L_2 \times L_3 \times L_4$$

where

$$L_1 = \frac{C_1}{\sigma^n} \exp\left[-\frac{1}{2\sigma^2} \sum_1^k (\underline{Y}_i - \underline{X}_i^t \beta_i)^t (\underline{Y}_i - \underline{X}_i^t \beta_i)\right]$$

$$L_2 = \frac{C_2}{\sigma} \exp\left[-\frac{1}{2\sigma^2} (Y^* - \sum_i \theta_i \underline{X}^{*t} \beta_i)^2\right]$$

$$L_3 = \frac{C_3}{|\,\underline{\Sigma}\,|^{n/2}} \cdot \exp\left[-\frac{1}{2} \sum_{i=1}^k \sum_{j=1}^{n_i} (\underline{X}_{ij} - \mu_i)^t \underline{\Sigma}^{-1} (\underline{X}_{ij} - \mu_i)\right]$$

$$L_4 = \frac{C_4}{|\,\underline{\Sigma}\,|^{1/2}} \exp\left[\frac{1}{2} \sum_{i=1}^k \theta_i (\underline{X}^* - \mu_i)^t \underline{\Sigma}^{-1} (\underline{X}^* - \mu_i)\right]$$

Taking logs:

$$\log L = P + Q + \text{constant}$$

where

$$P = -\frac{1}{2\sigma^2}\left[\sum_1^k (\underline{Y}_i - \underline{X}_i^t \beta_i)^t (\underline{Y}_i - \underline{X}_i^t \beta_i) + (Y^* - \sum_1^k \theta_i \underline{X}^{*t} \beta_i)^2\right] - \frac{n+1}{2} \log \sigma^2$$

$$Q = -\frac{n+1}{2}\log|\underline{\Sigma}|$$

$$-\frac{1}{2}\Big[\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\underline{X}_{ij}-\underline{\mu}_i)^t\,\underline{\Sigma}^{-1}(\underline{X}_{ij}-\underline{\mu}_i)$$

$$+\sum_{i=1}^{k}\theta_i(\underline{X}^*-\underline{\mu}_i)^t\,\underline{\Sigma}^{-1}(\underline{X}^*-\underline{\mu}_i)\Big]$$

In the next two sections we find the conditional maximum likelihood estimates (MLE) given $\theta_s = 1$. Given $\theta_s = 1$ denote the conditional maximum likelihood estimates of the parameters by $\hat{\underline{\beta}}_i$, $\hat{\underline{\mu}}_i$, $\hat{\underline{\Sigma}}$, $\hat{\sigma}^2$, $\hat{Y}^*$.

Conditional MLE of Y^*, σ^2 and $\underline{\beta}_i$

$\frac{\partial \log L}{\partial \underline{\beta}_s} = 0$ gives

$$(\underline{X}_s\underline{X}_s^t + \underline{X}^*\underline{X}^{*t})\,\hat{\underline{\beta}}_s = \underline{X}_s\underline{Y}_s + \underline{X}^*\hat{Y}^*. \qquad (1)$$

$\frac{\partial \log L}{\partial \underline{\beta}_i} = 0$ for $i \neq s$ gives

$$\underline{X}_i\underline{X}_i^t\hat{\underline{\beta}}_i = \underline{X}_i\underline{Y}_i. \qquad (2)$$

$\frac{\partial \log L}{\partial \sigma^2} = 0$ gives

$$\hat{\sigma}^2 = \frac{1}{n+1}\Big[\Sigma(\underline{Y}_i - \underline{X}_i^t\hat{\underline{\beta}}_i)^t(\underline{Y}_i - \underline{X}_i^t\hat{\underline{\beta}}_i) + (\hat{Y}^* - \underline{X}^{*t}\hat{\underline{\beta}}_s)^2\Big]. \qquad (3)$$

$\frac{\partial \log L}{\partial Y^*} = 0$ gives

$$\hat{Y}^* = \underline{X}^{*t}\hat{\underline{\beta}}_s. \qquad (4)$$

Substituting (4) in (1) we get

$$\underline{X}_s\underline{X}_s^t\hat{\underline{\beta}}_s = \underline{X}_s\underline{Y}_s.$$

Thus $\hat{\underline{\beta}}_i = (\underline{X}_i\underline{X}_i^t)^{-1}(\underline{X}_i\underline{Y}_i)$ for all $i = 1, \ldots, k$. (5)

Substituting (4) in (3) we get

$$\hat{\sigma}^2 = \frac{1}{n+1}\sum_i(\underline{Y}_i - \underline{X}_i^t\hat{\underline{\beta}}_i)^t(\underline{Y}_i - \underline{X}_i^t\hat{\underline{\beta}}_i). \qquad (6)$$

Conditional MLE of $\underline{\mu}_i$ and $\underline{\Sigma}$

In the expression for the likelihood above only the terms L_3 and L_4 contain $\underline{\mu}_i$ and $\underline{\Sigma}$. Thus we can find the MLE's of $\underline{\mu}_i$ and $\underline{\Sigma}$ by maximizing $L_3 \times L_4$. But since $\theta_s = 1$ we simply have a situation where there are n_i observations from $N(\underline{\mu}_i, \underline{\Sigma})$, $i \neq s$, and $n_s + 1$ from $N(\underline{\mu}_s, \underline{\Sigma})$.

Thus the conditional MLE's of $\underline{\mu}_i$ and $\underline{\Sigma}$ can be obtained from standard multivariate theory (e.g. Anderson, (1958) p. 248). These are

$$\hat{\underline{\Sigma}} = \frac{1}{n+1}\underline{A} = \frac{1}{n+1}\sum_{i=1}^{k}\underline{A}_i$$

where

$$\underline{A}_i = \sum_{j=1}^{n_i}(\underline{X}_{ij}-\bar{\underline{X}}_i)(\underline{X}_{ij}-\bar{\underline{X}}_i)^t, \text{ for } i \neq s$$

and

$$\underline{A}_s = \sum_{j=1}^{n_s}(\underline{X}_{sj}-\bar{\underline{X}}_s)(\underline{X}_{sj}-\bar{\underline{X}}_s)^t + (\underline{X}^*-\bar{\underline{X}}_s)(\underline{X}^*-\bar{\underline{X}}_s)^t$$

and finally

$$\hat{\underline{\mu}}_i = \bar{\underline{X}}_i.$$

Estimating $\underline{\theta}$

Substituting these conditional MLE's into the expression for the likelihood above we see that max $(L_1 \times L_2)$ is independent of s. The maximum value of $L_3 \times L_4$ is

$$\frac{C}{|\hat{\underline{\Sigma}}|^{\frac{n+1}{2}}}$$

where $\hat{\underline{\Sigma}}$ is a function of s, and so we will denote it by $\hat{\underline{\Sigma}}^{(s)}$. Thus the MLE of s is that value which minimizes $|\hat{\underline{\Sigma}}^{(s)}|$ or equivalently $|\underline{A}^{(s)}|$.

Discussion

1. The regression is estimated from $n\ (= \Sigma n_i)$ observations on which y and $\underline{x}$ are available.

2. Classification is based on all the $n + 1$ observations on which $\underline{x}$ is available.

3. The procedure operates as follows:

   a) Compute

   $$|\underline{A}^{(s)}| = \Big|\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\underline{X}_{ij}-\bar{\underline{X}}_i)(\underline{X}_{ij}-\bar{\underline{X}}_i)^t + (\underline{X}^*-\bar{\underline{X}}_s)(\underline{X}^*-\bar{\underline{X}}_s)^t\Big|$$

   for all $s = 1, \ldots, k$.

   b) If r is the value of s for which $|\underline{A}^{(s)}|$ is minimum, assign $\underline{X}^*$ to the r-th population.

c) Estimate Y^* by

$$\hat{Y}^* = \underline{X}^{*t}\,\hat{\beta}_r$$

where

$$\hat{\beta} = (\underline{X}_r\underline{X}_r^t)^{-1}(\underline{X}_r\underline{Y}_r)$$

Pooled Regressions

We assume the usual regression model in each population, and we let the true regression in the i-th population be denoted by R_i, the estimated (least squares) regression by $\hat{R}_i$, the variance of the dependent variable by η_i^2, and the standard error of estimate by σ_i. Thus

$$y_{ij} = R_i + \epsilon_{ij} \qquad j = 1, \ldots, n_i$$

where $\epsilon_{ij} \sim N(0, \eta_i^2)$

For the special case of simple linear regression,

$$R_i = \alpha_i + \beta_i x$$

$$\hat{R}_i = \hat{\alpha}_i + \hat{\beta}_i x$$

and

$$\sigma_i^2 = \eta_i^2\left(\frac{1}{n_i} + \frac{(x - \bar{x}_i)^2}{\sum_j (x_{ij} + \bar{x}_i)^2}\right)$$

where $\hat{\alpha}_i$ and $\hat{\beta}_i$ are the usual least squares estimates.

The population of origin of the new individual is unknown, but we suppose that the probability he came from the i-th population is π_i, $i = 1, \ldots, k$.

We propose to investigate several methods of pooling the regressions. Which method is best will depend on the criterion used to evaluate the predicting ability of these methods. We propose to use the mean square error (MSE) since this is attractive in itself, and can be looked at as comprising the variance of the estimator plus the square of the bias. Thus minimizing the MSE is controlling the size of both these parameters.

Expectation Estimator (EE)

The simplest pooled estimator is

$$EE = \Sigma\, \pi_i \hat{R}_i \tag{7}$$

The expected value of this is $\Sigma\, \pi_i R_i$. For convenience we shall denote this by $\bar{R}$. Thus

$$E(EE) = \Sigma\, \pi_i R_i = \bar{R} \tag{8}$$

The expected MSE in predicting the mean value of y is

$$MSE(EE) = \sum_j \pi_j \left[\sum_i \pi_i^2 \sigma_i^2 + (\bar{R} - R_j)^2\right]$$

$$= \Sigma\, \pi_i (R_i - \bar{R})^2 + \Sigma\, \pi_i^2 \sigma_i^2 \tag{9}$$

$$= \Sigma\, \pi_i R_i^2 - \bar{R}^2 + \Sigma\, \pi_i^2 \sigma_i^2$$

A Weighted Expectation Estimator (WEE)

Since we are using minimum MSE as our criterion of goodness we are not restricted to using estimates whose expected bias is zero. Thus we consider

$$WEE = a\, \Sigma\, \pi_i \hat{R}_i \tag{10}$$

where a is a constant to be determined. To find the value of a which minimizes the MSE we set equal to zero the derivative of

$$MSE = a^2 \sum_i \pi_i^2 \sigma_i^2 + \sum_j \pi_j (a\bar{R} - R_j)^2$$

Solving for a we get

$$a = \frac{\bar{R}^2}{\Sigma\, \pi_i^2 \sigma_i^2 + \bar{R}^2} \tag{11}$$

For this optimum value of a the expected value and the MSE of this estimate are

$$E(WEE) = \frac{\bar{R}^2}{\Sigma\, \pi_i^2 \sigma_i^2 + \bar{R}^2} \cdot \bar{R} \tag{12}$$

$$MSE(WEE) = \Sigma\, \pi_i (R_i - \bar{R})^2 + \frac{\bar{R}^2}{\Sigma\, \pi_i^2 \sigma_i^2 + \bar{R}^2} \cdot \Sigma\, \pi_i^2 \sigma_i^2 \tag{13}$$

Optimal Weighted Estimator (OWE)

The two estimators considered so far do not take into consideration the possibility that there may be different sample sizes and therefore differing amounts of information available in the different populations. So let us consider an estimate

$$OWE = \Sigma\, a_i \hat{R}_i \tag{14}$$

where the a_i, $i = 1, \ldots, k$ are constants chosen to minimize the expected MSE:

$$MSE = \sum_i a_i^2 \eta_i^2 + \sum_j \pi_j \left(\sum_i a_i R_i - R_j\right)^2$$

Differentiating this with respect to a_i, $i = 1, \ldots, k$, and setting the derivatives equal to zero we solve for the a_i:

$$a_i = \frac{\frac{R_i}{\sigma_i^2}\bar{R}}{1 + \Sigma \frac{R_i^2}{\sigma_i^2}} \qquad (15)$$

The estimator, the expected value, and the MSE using these optimal a_i's are

$$OWE = \frac{\bar{R}\,\Sigma \frac{R_i\hat{R}_i}{\sigma_i^2}}{1 + \Sigma \frac{R_i^2}{\sigma_i^2}} \qquad (16)$$

$$E(OWE) = \frac{\Sigma \frac{R_i^2}{\sigma_i^2}}{1 + \Sigma \frac{R_i^2}{\sigma_i^2}} \cdot \bar{R} \qquad (17)$$

$$MSE(OWE) = \Sigma\, \pi_i (R_i - \bar{R})^2 + \frac{\bar{R}^2}{1 + \Sigma \frac{R_i^2}{\sigma_i^2}} \qquad (18)$$

Constrained Weighted Estimator (CWE)

For those who would like the coefficients to sum to unity we offer:

$$CWE = \Sigma\, a_i \hat{R}_i \quad \text{where } \Sigma\, a_i = 1. \qquad (19)$$

To find the optimum a_i we differentiate with respect to a_i, $i = 1, \ldots, k$

$$\varphi = \Sigma\, a_i^2 \sigma_i^2 + \sum_j \pi_j (\sum_i a_i R_i - R_j)^2 + \lambda(\sum_i a_i - 1)$$

and set the derivatives equal to zero. To solve these equations we first multiply the i-th equation by $\frac{1}{\sigma_i^2}$ and add over i. Then multiply the i-th equation by $\frac{R_i}{\sigma_i^2}$ and add over i. These two equations can be solved for $\Sigma\, a_i R_i$ and λ, and thus we can solve for a_i:

$$a_i = \frac{1}{D}\left[\frac{1}{\sigma_i^2}(1 + \Sigma \frac{R_i^2}{\sigma_i^2} - \Sigma \frac{R_i}{\sigma_i^2}\bar{R}) + \frac{R_i}{\sigma_i^2}(\Sigma \frac{1}{\sigma_i^2}\bar{R} - \Sigma \frac{R_i}{\sigma_i^2})\right] \qquad (20)$$

where $D = (1 + \Sigma \frac{R_i^2}{\sigma_i^2})\, \Sigma \frac{1}{\sigma_i^2} - (\Sigma \frac{R_i}{\sigma_i^2})^2$ which is positive by Schwarz' inequality.

Thus the expected value and the MSE of the CWE are

$$E(CWE) = \bar{R} + \frac{1}{D}(\Sigma \frac{R_i}{\sigma_i^2} - \bar{R}\,\Sigma \frac{1}{\sigma_i^2}) \qquad (21)$$

$$MSE(CWE) = \Sigma\, \pi_i (R_i - \bar{R})^2 + \frac{\bar{R}^2}{1 + \frac{R_i^2}{\sigma_i^2}} + \frac{(1 + \Sigma \frac{R_i^2}{\sigma_i^2} - \bar{R}\,\Sigma \frac{R_i}{\sigma_i^2})^2}{D(1 + \Sigma \frac{R_i^2}{\sigma_i^2})} \qquad (22)$$

Lumped Estimator (LE)

Up to this point we have had to estimate the regression in each population, and then make our prediction using some set of weights. An alternative method is to simply lump all observations together as if they were from one large population, estimate the regression in that large population, and use this regression for prediction. We shall call this the Lumped Estimator (LE).

One situation where this would be a desirable alternative is where the degrees of freedom in the separate populations are rather small for the separate estimations.

Comparing the ability to predict of the LE with that of earlier weighted estimators becomes very involved. We present here a comparison for the case of simple linear regression, and we make one or two simplifying assumptions. We shall use the following notation:

$$LE = \hat{\alpha} + \hat{\beta}\, x \qquad (23)$$

$$\hat{\beta} = \frac{\Sigma\, \Sigma\, (y_{ij} - \bar{y})(x_{ij} - \bar{x})}{D_x}$$

$$D_x = \Sigma\, \Sigma\, (x_{ij} - \bar{x})^2 = n(W + B) \qquad (24)$$

$$W = \frac{1}{n}\Sigma\, \Sigma\, (x_{ij} - \bar{x}_i)^2$$

$$B = \Sigma\, p_i (\bar{x}_i - \bar{x})^2$$

$$p_i = \frac{n_i}{n} \qquad (25)$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x}$$

$$\bar{y} = \Sigma\, p_i \bar{y}_i$$

$$\bar{x} = \Sigma\, p_i \bar{x}_i$$

Now it is easily shown that

$$\hat{\beta} = \frac{\Sigma\, D_i \hat{\beta}_i + \Sigma\, n_i (\bar{x}_i - \bar{x})\bar{y}_i}{D_x}$$

where

$$D_i = \Sigma(x_{ij} - \bar{x}_i)^2$$

For simplification we shall assume that it is approximately true that

$$\frac{D_i}{n_i} = W \quad \text{for all } i = 1, \ldots, k.$$

Note that this is assuming that the "within-sample" variance of the x_{ij} is approximately the same in all populations. In this case

$$\hat{\beta} = \frac{W}{W + B} \Sigma p_i \hat{\beta}_i + \frac{B}{W + B} \hat{\beta}^*$$

where
$$\hat{\beta}^* = \frac{\Sigma p_i(\bar{x}_i - \bar{x})\bar{y}_i}{\Sigma p_i(\bar{x}_i - \bar{x})^2} \qquad (26)$$

Note that $\hat{\beta}^*$ is the slope of a weighted regression line fitted to the centroids of the k samples.

Thus the lumped estimator is

$$\begin{aligned} LE &= \bar{y} + (x - \bar{x}) \Sigma p_i\hat{\beta}_i \\ &\quad + \frac{B}{W + B} (x - \bar{x})(\hat{\beta}^* - \Sigma p_i\hat{\beta}_i) \\ &= \Sigma p_i\hat{R}_i + \Sigma p_i\hat{\beta}_i(\bar{x}_i - \bar{x}) \\ &\quad + \frac{B}{W + B} (x - \bar{x})(\hat{\beta}^* - \Sigma p_i\hat{\beta}_i) \qquad (27) \end{aligned}$$

The expected value is

$$\begin{aligned} E(LE) &= \Sigma p_iR_i + \Sigma p_i\beta_i(\bar{x}_i - \bar{x}) \\ &\quad + \frac{B}{W + B} (x - \bar{x})(\beta^* - \bar{\beta}) \qquad (28) \end{aligned}$$

where $\bar{\beta} = \Sigma p_i\beta_i$ (29)

Note that β^* may be regarded as the slope of the regression between populations, and $\bar{\beta}$ an average slope within populations.

Now LE is the usual least squares estimator. Let us assume for the sake of simplification that the variance of y in the i-th population is a constant, i.e.

$$\eta_i = \eta, \quad \text{for all } i = 1, \ldots, k. \qquad (30)$$

Thus

$$\text{var}(LE) = \frac{\eta^2}{n} \left(1 + \frac{(x - \bar{x})^2}{W + B}\right) \qquad (31)$$

Note that if the η_i are not assumed equal the variance of LE can be found from an alternative form of LE, namely

$$\begin{aligned} LE = \Sigma p_i\Big\{ &\left[1 + \frac{(x - \bar{x})(\bar{x}_i - \bar{x})}{W + B}\right] \bar{y}_i \\ &+ \frac{W}{W + B} (x - \bar{x}) \hat{\beta}_i\Big\} . \end{aligned}$$

The MSE of the Lumped Estimator

If, as earlier, we let π_i denote the probability that the individual comes from the i-th population we can write down the expected MSE:

$$\text{MSE}(LE) = \text{var}(LE) + \Sigma \pi_i(R_i - E(LE))^2$$

where var(LE) and E(LE) are given in equations (31) and (28).

There now arises the problem of comparing this MSE with the MSE of earlier pooled estimators. This is quite difficult, especially if we can assume nothing about the relationship between p_i and π_i. Even if the total sample is random, a complete analysis would discuss the variance of p_i as and estimate of π_i. If the sample is large p_i would approximate π_i. But anyway, in order that we can get a feel for the relative goodness of LE as opposed to EE, the expectation estimator, we shall assume that it is approximately true that

$$p_i = \pi_i \qquad (32)$$

In this case we can show that

$$\begin{aligned} \text{MSE}(LE) &= \frac{\eta^2}{n} \left(1 + \frac{(x - \bar{x})^2}{W + B}\right) + \Sigma \pi_i(R_i - \bar{R})^2 \\ &\quad + \Big[\Sigma \pi_i\beta_i(\bar{x}_i - \bar{x}) \\ &\quad + \frac{B}{W + B} (x - \bar{x})(\beta^* - \bar{\beta})\Big]^2 \qquad (33) \end{aligned}$$

For comparing this with MSE(EE) in equation (9) we note first that σ_i^2 can be written:

$$\sigma_i^2 = \frac{\eta_i^2}{n_i} \left(1 + \frac{(x - \bar{x}_i)^2}{W}\right),$$

and since we are assuming $\eta_i = \eta$ for all i, and that $n_i = n\,\pi_i$,

$$\begin{aligned} \Sigma \pi_i^2\sigma_i^2 &= \frac{\eta^2}{n} \Sigma \pi_i\left(1 + \frac{(x - \bar{x}_i)^2}{W}\right) \\ &= \frac{\eta^2}{n} \left(1 + \frac{(x - \bar{x})^2}{W} + \frac{B}{W}\right) \end{aligned}$$

Thus from equation (9)

$$\begin{aligned} \text{MSE}(EE) &= \frac{\eta^2}{n} \left(1 + \frac{(x - \bar{x})^2}{W} + \frac{B}{W}\right) \\ &\quad + \Sigma \pi_i(R_i - \bar{R})^2. \qquad (34) \end{aligned}$$

Comparison of the Pooled Estimators

It can be seen from equations (9), (13), (18), and (22) under what circumstances the various pooled estimators would be superior, and by how much. It is clear that OWE is best, but it must be remembered that the optimal weights are functions of the parameters, and therefore have to be estimated. Estimating these weights will obviously increase the MSE of these estimators. Since it seems reasonable that any improvement, if any, would be due to a large extent to an improved estimate of the variance, an investigation of the MSE of the estimated OWE should estimate the R_i's and the σ_i's.

One observation on MSE(OWE). We note that this is small whenever at least one coefficient of variation $\frac{R_i}{\sigma_i}$ is large at the value of x we are estimating for. A small σ_i provides some evidence that the new individual came from the i-th population. And from equation (15), if $\frac{R_i}{\sigma_i^2}$ is large then the weight for the i-th is large.

In an unpublished paper Francis (1969) considers the related problem of estimating the regression in a particular population when sample observations are also available from other populations which are considered somewhat similar. There again, a comparison of the pooled estimators will have to investigate the effect of estimating the optimal weights.

Comparison of EE and LE

By comparing equations (33) and (34) we can say something about EE versus LE. For example, suppose

$$\Sigma\, \pi_i \beta_i (\bar{x}_i - \bar{x}) = 0$$

This would be approximately true, for instance, if all the β_i are roughly equal, or alternatively if all the $\bar{x}_i$ were roughly equal. (Figures 1 and 2 display the concentration ellipses for two examples for which k = 3, where all the β_i are roughly equal, and where the within-sample variances of x would be approximately all equal. The slope of the line AB is $\bar{\beta}$ and the slope of CD is β^*.) In this case the difference between (33) and (34) is

$$\text{MSE(LE)} - \text{MSE(EE)} = \frac{T^2}{n} \left\{ z^2\left[\left(1 + \frac{B}{W}\right)^{-1} - 1\right] - \frac{B}{W}\right\} + \frac{B^2W}{(W+B)^2}\, z^2(\beta^* - \bar{\beta})^2 \qquad (35)$$

where $$z^2 = \frac{(x - \bar{x})^2}{W}.$$

The following observations can be made:

i) Very small B could imply a very large β^*: EE is better (see Figure 2).

ii) When B approaches W in size and when $(\beta^* - \bar{\beta})^2 \gg 0$, then EE is better, except when $x \doteq \bar{x}$ and then:

iii) When B approaches W in size, and $x \doteq \bar{x}$ then LE is best.

iv) When B is extremely large we would surely modify our π_i's, given the x observation on our new individual, to eliminate most populations, unless the original samples were not random. (See section on Posterior Probabilities below.)

Posterior Probabilities

We have been assuming that π_i is known for all i. But as was suggested in observation (iv) above, if these are only prior probabilities, and if the original observations on (x_i) were a random sample from all populations, then these sample observations, together with the observations on (x_i) for the new individual, should be used to modify the π_i's. We shall not pursue this topic in this paper.

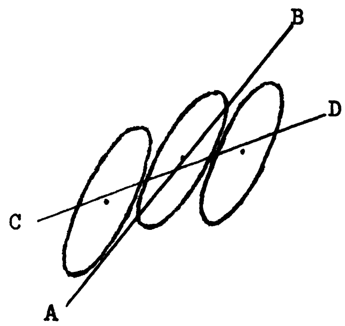


Figure 1

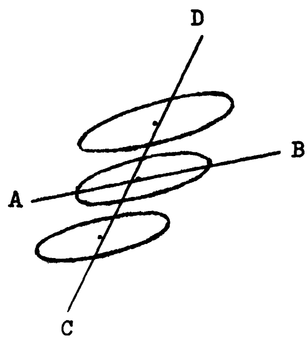


Figure 2

References

Anderson, T. W., Introduction to Multivariate Statistical Analysis. New York: John Wiley and Sons, Inc., 1958.

Francis, I. S., "On Pooling Regressions," Unpublished, 1969.

Hartley, H. O. and Rao, J. N. K., "Classification and Estimation in Analysis of Variance Problems," Review of the International Statistical Institute, Vol. 36: 2, 1968, pp. 141-147.

XIV

CONTRIBUTED PAPERS IV

NEW APPROACHES FOR SURVEY DESIGN AND RESPONSE PROBLEMS

Chairman, ELIJAH WHITE, National Center for Health Statistics

A TECHNIQUE FOR REDUCING SURVEY SAMPLING SIZE*

R. Hoffman**, E. Gomolka***, and H. Phillips****

> What can we do when the survey must produce several, perhaps many, variates? This calls for caution before resorting to disproportionate sampling... If the different answers point toward the same design, there is no great problem. But if the answers are contradictory, the choice involves good judgment and theory that is beyond existing theoretical developments.[1]

Introduction

This paper briefly describes the increasing need for the use of survey type sampling in the development and evaluation of action research projects. The standard methods appear to have some limitations when one is faced with stratum or cells within the social survey which have a finite (often very small) number of possible observations and/or surveys where there are a large number of attributes which one may wish to sample.

A method and accompanying computer based procedure is presented. The procedure specifies the sample size to be taken for each cell without regard to whether the particular cell is sampled.

It is shown that the method presented results in a lower total sample than fixed sample size methods and that the method converges to fixed sample size when the number of possible observations in each cell becomes very large. Finally, an extension to the method using an economical pilot sample, and the basis for the selection of acceptable sample size for the cells which are used, is discussed as a corollarly result.

The Problem

The objective of the social survey is primarily to describe a population in order to make inferences relating to phenomena which may or may not exist within the population. Typically, social survey techniques have made use of base data, such as census data, which are regularly collected by Federal agencies. Often, however, in action-oriented research programs conducted by the government and universities, the data required to evaluate these programs is very rarely available from the standard sources. Thus, the researcher must survey the target population for the data needed to evaluate the program. Similar problems often occur when attempting to obtain information with which to develop new programs. Hence, in evaluating or developing new government programs, the researcher must often collect his own data. But with the increased need for social survey sampling, greater attention will be required for the development of sampling techniques which deal with

*The research of the first author was supported in part by a Public Health Service grant to the Center for Planning and Development Research, and a Department of Transportation grant (Cal. MTD-11) to the School of Criminology, both of the University of California at Berkeley. The authors received additional support from the School of Management, State University of New York at Buffalo.

**Department of Management Science, State University of New York at Buffalo and Visiting, School of Criminology, University of California at Berkeley.

***School of Management, State University of New York at Buffalo.

**** Department of Management Science, State University of New York at Buffalo.

1. Kish, Leslie, Survey Sampling, New York (Wiley), 1967, pp.96-97.

2. See, P.G. Gray and T. Corlett, "Sampling for the Social Survey," Journal of the Royal Statistical Society, Part II, pp.150-199 (1950), for a basic discussion of multi-stage procedures for social survey sampling when the survey is to be used as an alternative to, or to update, a census. Similar procedures are appropriate for the problems we discuss.

cases where the usual assumptions are not appropriate.[3]

Two cases where significant problems may arise with the use of the usual techniques are: 1) a finite number of possible observations within a cell or stratum to be sampled, and 2) sampling of populations to obtain information on relatively rare events. We find many of the cells within the population may have a small number of possible observations. For example, if we are evaluating the effects of an existing poverty program it may be useful to stratify on the basis of head of household yearly income. But it is likely that in very affluent suburbs we may find in these strata there are very few individuals with incomes below $3000. Alternatively, a converse situation will hold in ghetto neighborhoods. However, it is often very important to obtain information on these unusual situations. A second source of finiteness is that the events of interest may be relatively rare, e.g., murder and rape in a given precinct in a given city. Hence, the cost of sampling (including the questionnaire cost and the cost of transmitting the interviewer to the area (load cost) becomes a major problem, when one considers the number of interviews which are often necessary to obtain a relevant sample.

The methods are, of course, quite standard if one wishes to obtain data from the population on but one attribute. However, this is not the typical case. If we survey a large population for one attribute, the value of this information will most often not be sufficient to justify the costs of sampling. Additionally, one may not know a priori those attributes which will be of importance.[4] Hence, the "optimal method" of Neyman as described in Mills[5] results in difficulty since, with its use, sample size will be functionally related to the variance of but one attribute.

However, a prevalent case upon which we focus our attention has a finite (and many times very small) number of possible observations in many of the strata or cells. When this is the case, proportional sampling implies that one may sample say 40 of a population of 400 houses in one subdivision but 1 of a population of 8 houses in another subdivision. Clearly, the sample of 40 may give us very acceptable results (if we wish to obtain inferences relating to the subdivision containing 400 homes), but the sample of 1 out of 8 very likely will not.[6]

The obvious alternative is to use a fixed sample size. When this is done, we may find that the appropriate sample size, say 30, may leave us with a sample size greater than population size.[7] Clearly, the information obtained from a complete enumeration of the attributes of the population will will exceed that obtained from any sample of the population.[8] Complete enumeration of some cells results in oversampling of such cells relative to

---

3. See, D. Raj, Sampling Theory, New York: McGraw-Hill, (1968), p.36, for a standard discussion. However, when there are but a finite possible number of observations the sampling fraction n_i/N_i is not small, hence the assumptions relating to the dependence of the variance of mean upon n (sample size) and σ^2 (variance) are not appropriate.

4. As a collection of 300 attributes or more is not unusual, it will be necessary to obtain information on many attributes to determine the crucial variables.

5. See, F.C. Mills, Statistical Methods, 3rd Ed., New York: Holt, Rinehart, and Winston, (1955), p.180. The other alternatives proportional and fixed sample size, can be used.

6. The arguments in favor of the proportional sampling method have often been based upon ease of computation; e.g., Kish, op.cit., Chapter 3. However, as an appropriate weighting scheme can always be found and with the availability of high-speed computing equipment, the extra costs of weighting and special handling are minimal.

7. See previous example for such a case: e.g., total subdivision population of 8 houses.

8. Raj, op.cit., p.27, implies that complete enumeration is appropriate if each cell is to be sampled (i.e., no savings in load costs) but good sampling procedure will still result in lower costs with little if any loss of precision. Also, using standard techniques such as those described in Raj, op.cit., p.81 and p.206, it may not be necessary to sample all cells.

those cells in which but a fraction are taken. Hence, when one has a finite number of possible observations in a cell, if a method can be found which results in equal precision and reliability for each cell, it is possible to reduce the sample size in the cells with a finite possible number of observations. This, of course, would result in lowered sampling costs.

The Method and Computer Based Procedure

The method for the selection of the sample size for each cell can be developed using the basic homoscedasticity assumption.

This method is derived from the basic equations for the standard error.

Here S_x^2 is the unbiased estimator of σ_x^2. Then we obtain n(sample size) from the standard relationship between precision and reliability.

$$e^2 = z^2\sigma_x^2$$ [9]

Substituting s_x^2 for σ_x^2

$$e^2 = z^2 s_x^2 = z^2 \frac{s^2}{n}$$

Solving for N and applying the exact form of the finite population correction factor[10] (fpc), as it is appropriate in the cases of interest.

$$n = \frac{z^2x^2}{e^2} \cdot \left(1 - \frac{n-1}{N-1}\right)$$

We define $C_1 = \frac{s^2}{e^2}$, where we accept the standard homoscedasticity assumption.

Then, for populations with a finite number of observations, the "t" distribution is utilized. This is necessary, as quite often the sample size will be small.

9. T. Yamane, Elementary Sampling Theory, Englewood Cliffs, N.J.: PrenticeHall, Inc., 1967, p.81.

10. Raj, op.cit., p.36, 1- = fpc

$$n = C_1t^2\left(1 - \frac{(n-1)}{(N-1)}\right)$$

$$n(N-1) = C_1t^2(N-1) - C_1t^2(n-1)$$

$$n(N-1+C_1t^2) = C_1t^2(N-1) + C_1t^2$$

$$n = \frac{C_1t^2+N}{C_1t^2+(N-1)}$$

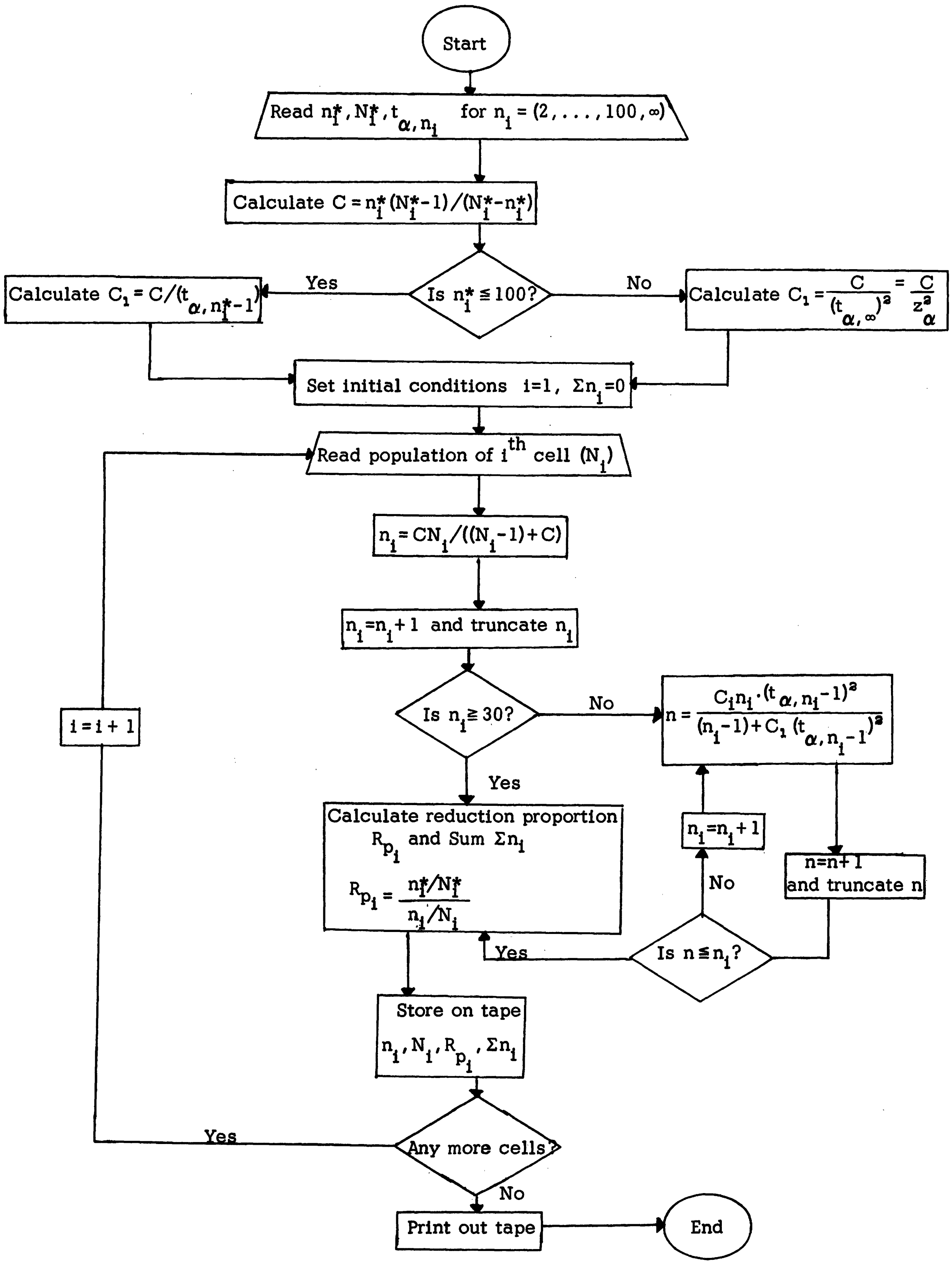
if n sufficiently large then $t^2 \cong z^2$

and $C_1z^2 = e$

$$n = \frac{CN}{C+(N-1)}$$

If the researcher is willing to specify an acceptable sample size (n_i^*) for any cell, then given the N_i (population size) for all other cells it is possible to specify the n_i for all other cells. The method uses 1) n_i^*, 2) N_i^* (population size) of cell i*, and 3) a range of values from the "t" distribution for "t" at a researcher specified level of alpha (α).

The computer based procedure is described in the accompanying flow chart (on the following page).



Start
Read n*ᵢ, N*ᵢ, t α,nᵢ for nᵢ = (2, . . . , 100, ∞)
Calculate C = n*ᵢ (N*ᵢ-1)/(N*ᵢ-n*ᵢ)
Is n*ᵢ ≦ 100?
Yes
Calculate C₁ = C/(t α,n*ᵢ-1)
No
Calculate C₁ = C/(t α,∞)² = C/z²α
Set initial conditions i=1, Σnᵢ=0
Read population of ith cell (Nᵢ)
nᵢ = CNᵢ/((Nᵢ-1)+C)
nᵢ=nᵢ+1 and truncate nᵢ
Is nᵢ≧30?
No
n = C₁nᵢ·(t α,nᵢ-1)² / (nᵢ-1)+C₁ (t α,nᵢ-1)²
Yes
Calculate reduction proportion
Rpᵢ and Sum Σnᵢ
Rpᵢ = (n*ᵢ/N*ᵢ) / (nᵢ/Nᵢ)
nᵢ=nᵢ+1
n=n+1
and truncate n
No
Yes
Is n≦nᵢ?
Store on tape
nᵢ, Nᵢ, Rpᵢ, Σnᵢ
i=i+1
Yes
Any more cells?
No
Print out tape
End

Conclusion and Extension

The method can be described simply. If one is willing to accept a sample size (n*) for any cell in a stratified and/or multi-stage sample, then with the standard homoscedasticity assumptions, we can specify the sample size for all cells which will result in a total sample size which is usually less, and never greater, than the total sample size obtained from a fixed sample size procedure of equivalent effectiveness.

It can be shown where:

$N_i \to \infty$ (actually very large N_i)

then

$n_i \to n_f$ (fixed sample size)[11]

since

$$n_i = \frac{C \cdot N_i}{(N_i - 1) + C}$$

the one can show as $N \to \infty$ (or is very large)

$$n_i = \frac{C \cdot N_i}{(N_i - 1) + C} \to \frac{C \cdot N_i}{N_i + C} \to \frac{C \cdot N_i}{N_i} \to C = n_f$$

$n_i \cong C$ and a constant, i.e., fixed sample size.

A very interesting extension relates to the use of a pilot sample, say in a cell with relatively low load costs.[12] If on the basis of the pilot cell sample's variances one is willing to specify a sample size which would be acceptable for the pilot cell, then it is possible to specify the sample size for all cells which could be in the grand sample. This result only requires that one again accept the standard homoscedasticity assumption. It should be noted that it is not necessary that the pilot sample be included in the final grand sample.[13]

11. Obviously there exists a fixed sample size which would be acceptable.

12. Perhaps, in a nearby jurisdiction, prison, or housing subdivision.

13. This will occur rather often in practice, as the probability of the pilot cell being in the final strata or cell is a function of the ratio of the pilot population size to the size of the grand population,

Pr(pilot sample in grand sample)

$$= \frac{n_i}{N} \cdot \frac{N_i\,(\text{pilot})}{N}$$

THE OLDER PERSON AS A SURVEY RESPONDENT

Lola M. Irelan

Social Security Administration

Replies to the queries of a survey interviewer are not always factual--even when the respondent intends them to be. Such error may arise from multiple sources--from personality, from differential group membership, from reaction to an interviewer, from the influence of social status, etc. Allowing for such bias is a difficult task of uncertain outcome. It can be eased when systematic knowledge is available of the probable response of different categories of people to the interviewing situation itself. In the course of preparing to study a sample of older people for 10 years, and simultaneously to cope with their changing characteristics, a library search was made for information on aging effects which might have implications for the interview process. The results, which I am going to present to you, are a small store of apparently relevant facts and a few general hypotheses which are candidates for refinement and testing.

Three aspects of aging have clear implications for performance in a survey interview situation: (1) changes in psycho-physical abilities; (2) changes in response style; and (3) changes in the nature of values and motivation.

Most people are at least impressionistically aware of the psycho-physical changes brought about by aging. General visual acuity usually begins to decline at between 40 and 50 years of age. Among these 65 and over, 40 percent of the men and 60 percent of women have 20/70 acuity or less. 1/ In addition to this general decrease in optical acuteness, some more qualitative decrements set in. Judgment of dimension and direction is less accurate. Depth perception, judgment of line length, and assessment of verticality are less dependable than among the young. 2/ Discernment of embedded, incomplete, or ambiguous forms decreases. 3/ Increasing age, for most people, means marked loss in hearing, especially in ability to hear higher tones. Ability to discriminate speech also deteriorates. 4/ Decrement in this particular ability is of obvious importance in an interview situation. There is a loss with age of some sensorimotor coordination. 5/ Longer time is required to interpret, integrate, and act upon complex sets of stimuli.

The oldster's faltering memory is a byword. Considerable evidence supports it as objective fact. 6/ Both for memory over a short period of time and for long term recall the elderly are noticeably poorer performers than the young. Where the task of remembering is complicated (e.g., by using complex material, by introducing distractions, etc.) the difference is even more notable.

Some of the research on age related IQ change is equivocal. Nevertheless, it does seem clear that after about age 60, overall performance on IQ tests declines. 7/ Variously attributed to loss of sensory acuity, lowered response speed, and declining health--however caused, the change is a significant one for the interviewing process.

These losses are probably important sources of some age-related changes in individual response style. One notes his own increasing slowness, his dulling memory, his lessening ability to control the environment--and one tries out and probably adopts different techniques of responding to that environment. The most immediate shift is to decreased self-confidence. Older people consistently produce more negative, belittling self-concepts, than do middle-aged and younger respondents. 8/ Knowledge of one's increasing vulnerability probably brings about the typical caution and restraint of the elderly, and their more rigid approach to most situations.

The evidence for these age changes makes clear their relevance in the survey situation. The cautious, even wary, approach of the older person to a new situation--or to one requiring a decision--is a piece of popular lore which is well supported by research. Performance studies have several times shown older persons more willing to sacrifice speed for accuracy, and more likely to suppress natural responses. 9/ In other research, older people are found to respond more cautiously to hypothetical high risk, high reward situations. 10/ Underlying this increased caution is the suggestion of a generally more serious and deliberate attitude. 11/ They are less impulsive and more inclined to deliberate before making a decision. They indulge in more reviewing and mulling over of answers. 12/

In both attitudinal and behavioral research, the greater, rigidity of oldsters is clear. They are more dogmatic, less susceptible to pressure for the change of an attitude, and more likely to show sequential response patterns. They change less readily from one problem solving situation to another and perceive ambiguities less readily. 13/

Whatever the evidence on "disengagement"--and it seems to be inconsistent--it does seem clear that the quality of social involvement changes with age. And it does change in the direction of more detachment. Older people focus more upon themselves and their own emotions. Age, in general, is likely to mean more self-centeredness and less emotional responsiveness to the rest of the world. 14/

A natural outcome of physical and psychological aging is a shift in the hierarchy of personal values. The 60-year old has, consciously or not, done considerable testing of himself and the world around him--on both material and nonmaterial levels. He very likely has weighed his own experience of reality against what is socially held to be true and valuable. And he often comes to conclusions different from those of younger, inexperienced persons. Generally, it may be said that advanced age brings with it different personality needs and different topics of concern.

There are changes in three important need areas--need for achievement, need for power, need for affiliation. Older people are considerably less interested in culturally defined "achievement" than are the young. 15/ Interest in job promotion, for example, is lower. Conversely, the need for power over other people climbs slightly. 16/ Interest in extensive social interaction drops off.

There are marked shifts, with age, in degrees of interest or concern for certain topics, health and physiological concerns, financial matters, the topic of death (particularly one's own), and religion and morality. While generally conscious of physical well-being 17/ older people are more interested in their own health than are the young and middle-aged. They do more thinking about it as well as more talking with relatives and friends. Expectably enough, they also report more concern with physiological functioning. 18/ Financial matters are less worrisome to most older people. They think less often about their own financial situations, report less worry on that score, and are less likely to think they would be happier with more money. 19/ For most people, advancing age brings with it more conscious thinking about death. 20/ The prevailing attitude is not fearful, but rather accepting and un-tragic. 21/ Interest in religion seems to increase with age, 22/ as does evaluation of morality as a desirable characteristic. 23/

The more obvious hypotheses which might be based upon these known differentials between younger and older persons are proved out in existing reports. For example, several points would lead one to expect what is actually the case--that older people are more reluctant survey respondents. Their lowered persuasibility, lessened susceptibility to social pressure, and greater self-centeredness would make them tougher targets for any interviewer. In addition, the older person who is conscious of faltering memory and declining mental ability can be expected to shy away from questioning by strangers, especially when his answers are recorded. It is not surprising, then, that they are reported both to require more coaxing to grant interviews and to refuse more often. 24/ Similarly, they are less cooperative respondents to mail questionnaires. 25/

Failing memory and lowered ability to concentrate would logically make for less consistency from an initial interview to a later one, on the part of the aged. And indeed that seems to be the case as appears in at least one report. 26/ It would be interesting to do a study of this sort with the added element of either reminding or not reminding respondents of their earlier answers. The age-borne tendency to rigidity and to maintenance of attitudes should lead to greater consistency where older people can recall or be reminded of their previous statements.

The supposed "disengagement" of the aged has been suggested as a source of two types of respondent behavior. Whether or not disengagement is the best explanation, Gergen and Back 27/ have reported that older people tend to give more "no opinion" responses as well as more extreme responses--both events supposedly the result of resistance to interviewer attempts to cajole more refined answers from their sample members.

On the basis of the characteristics of the aged which have been summarized here, three additional general hypotheses may be proposed: (1) other things being equal, older people will require more time for an interview; (2) they will respond differentially to several interview topics; (3) they will react differently to interviewer characteristics.

Longer interview time should result from several things. Oldsters are more reluctant respondents and more time is probably used in persuading them to answer problem questions. Their nervous and intellectual reflexes are slower, they are more cautious and deliberate in decision making, so they require more time to select answers. Consciousness of failing memory and mental ability will also retard the speed of answers.

Changed interests and values should make for greater cooperation, on some topics, from older than from younger persons. They could be expected, for example, to respond more fully, and give more refined answers, to questioning on health matters. They will probably be more cooperative and straight forward in answering questions on their own income. Similarly, they will be more interested and responsive to topics relating to religion. Changing values will give different emotional tones to many specific topics--thereby differentiating older from younger respondents. For example, oldsters could be expected to be more detached in the reporting of their own occupational successes and failures--or in speculating about things related to death.

Several age-related changes make it reasonable to hypothesize that some interviewer characteristics will differentially affect older respondents. They should be more responsive to interviewers who are able to be fairly neutral stimulants, thereby offering the least occasion for any kind of a distracting power struggle between interviewer and respondent. There is likely to be a great difference in the quality of answers given to interviewers who speak distinctly and who are able--or permitted--to use explanatory transitions from one topic to another--more so than with younger people. They will probably be more cooperative with detached--and less with friendlier-interviewers than will younger people.

And so on. I have given you a far from exhaustive list of the effects of aging which may be important in conducting surveys and interpreting data there from. Plus a few examples of ways in which it seems the interview process might be affected.

FOOTNOTES

1/ National Center for Health Statistics, Vital and Health Statistics, 1964, Series 11, No. 3.

Weiss, Alfred D., "Sensory Functions," in Birren (ed), Handbook of Aging and the Individual, Chicago: University of Chicago Press, 1959, pages 503-542.

2/ Brebner, J., "The Effect of Age on Judgments of Length," Journal of Gerontology, 17 (1962) 91-94.

Hoffman, C.S., et al., "Effect of Age and Brain Damage on Depth Perception," Perceptual and Motor Skills, 9 (1959) 283-286.

Comalli, Peter E., Jr., "Cognitive Functioning in a Group of 80-90 Year Old Men," Journal of Gerontology, 20 (1965) 14-17.

3/ Basowitz, H., and S.J. Korchin, "Age Differences in the Perception of Closure," Journal of Abnormal and Social Psychology, 54 (1957) 93-97.

Axelrod, Seymour and Louis D. Cohen, "Senescence and Embedded Figure Performance in Vision and Touch," Perceptual Motor Skills, 12 (1961) 283-288.

Thomas, James M., and Don C. Charles, "Effects of Age and Stimulus Size on Perception," Journal of Gerontology, 19 (1964) 447-450.

4/ Melrose, J., Oliver L. Welsh, and David Luterman, "Auditory Responses in Selected Elderly Men," Journal of Gerontology, 18 (1963) 267-270.

Welsh, Oliver L., David M. Luterman, and Benjamin Bell, "The Effects of Aging on Responses to Filtered Speech," Journal of Gerontology, 24 (1969) 189-192.

5/ Riley, Matilda W., and Anne Foner, Aging and Society, Vol. I. New York: Russell Sage Foundation, 1968, pages 250-254.

Welford, Alan J., "Psychomotor Performance," in Birren (ed.), Handbook of Aging and the Individual, Chicago: University of Chicago Press, 1959, pages 562-613.

6/ Beard, Belle B., "Some Characteristics of Recent Memory of Centenarians," Journal of Gerontology, 23 (1968) 23-30.

Hultsch, David F., Rolf H. Monge, and Eric F. Gardner, "Adult Age Differences in Memory Performance and Memory Organization," paper read at Gerontological Society Meetings, 1968, Denver, Colorado.

Peak, Daniel J., "Changes in Short-Term Memory in a Group of Aging Community Residents," Journal of Gerontology, 23 (1968) 9-16.

Riley and Foner, op. cit., page 258.

Schonfield, David, and Wayne Donaldson, "Immediate Memory as a Function of Intraseries Variation," Canadian Journal of Psychology/Review of Canadian Psychology, 20 (1966) 218-227.

7/ Fozard, James L., "Age Related Changes in Scores on the General Aptitude Test Battery," paper read at Gerontological Society Meetings, 1968, Denver, Colorado.

Bromley, D.B., The Psychology of Human Aging, Baltimore: Penguin Books, 1966, pages 229-252.

Jarvik, Lissy F., and A. Folek, "Intellectual Stability and Survival in the Aged," Journal of Gerontology, 18 (1963) 173-176.

Jones, Harold E., "Intelligence and Problem-Solving," in Birren (ed.), Handbook of Aging and the Individual, Chicago: University of Chicago Press, 1959, pages 700-738.

8/ Kuhlen, Raymond G., "Personality Change with Age," in Philip Worchel and Dona Byrne (eds), Personality Change, New York: Wiley and Sons, Inc., 1964, pages 524-555.

9/ Riley and Foner, op. cit., 277, 278.

10/ Botwinick, Jack, "Cautiousness in Advanced Age," Journal of Gerontology, 21 (1966) 347-353.

Wallach, M.A., and N. Kogan, "Aspects of Judgment and Decision Making: Interrelationships and Changes with Age," Behavioral Scientist, 6 (1961) 23-26.

11/ Korchin, S.J., and H. Basowitz, "Age Differences in Verbal Learning," Journal of Abnormal and Social Psychology, 54 (1967) 64-69.

Eisdorfer, C., S. Axelrod and Frances L. Willkie, "Stimulus Exposure Time as a Factor in Serial Learning in an Aged Sample," Journal of Abnormal and Social Psychology, 67 (1963) 594-600.

12/ Botwinick, J., J.F. Brinley, and J.S. Robbin, "The Interaction Effects of Perceptual Difficulty and Stimulus Exposure Time on Age Differences in Speed and Accuracy of Response," Gerontologia, 2 (1958) 1-10.

13/ Taylor, C., "Adult Age Differences in Rigidity as Revealed in Attitude Scale Responses," American Psychology, 10 (1955) 339-340.

Shukin, Alexey, and Bernice L. Neugarten, "Personality and Social Interaction," in B.L. Neugarten, et al., Personality in Middle and Late Life, New York: Atherton Press, 1964, pp. 149-157.

14/ Dean, Lois R., "Aging and the Decline of Affect," Journal of Gerontology, 17 (1962) 440-446.

Gray, H., "Psychological Types and Changes with Age," Journal of Clinical Psychology, 3 (1947) 273-277.

Lakin, M., and C. Eisdorfer, "A Study of Affective Expression Among the Aged," in Tibbitts and Donahue (eds.), Social and Psychological Aspects of Aging, New York: Columbia University Press, 1962, pp. 650-654.

15/ Veroff, J., J.W. Atkinson, S.C. Feld, and G. Gurin, "The Use of Thematic Apperception to Assess Motivation in a Nationwide Interview Study," Psychological Monographs, 74, Number 499, 1960.

Neugarten, Bernice L., and D.L. Gutmann, "Age-Sex Roles and Personality in Middle Age: A Thematic Apperception Study," in B.L. Neugarten, et al., Personality in Middle and Late Life, New York: Atherton Press, 1964, pp. 44-89.

16/ Veroff, Atkinson, Feld and Gurin, op. cit. Neugarten and Gutmann, op. cit.

17/ Maddox, George L., "Some Correlates of Difference in Self-Assessment of Health Status Among the Elderly," Journal of Gerontology, 17 (1962) 180-185.

Suchman, Edward A., Bernard S. Phillips, and Gordon F. Streib, "An Analysis of the Validity of Health Questionnaires," Social Forces, 36 (1958) 223-232.

18/ Chown, Sheila M., and Alastair Heron, "Psychological Aspects of Aging in Man," in P. Farnsworth, et al. (eds), Annual Review of Psychology, Palo Alto: Annual Reviews, Inc., 1965, pp. 417-450.

19/ Back, Kurt W., and Kenneth J. Gergen, "Personal Orientation and Morale of the Aged," in Simpson and McKinney (eds.), Social Aspects of Aging, Durham: Duke University Press, 1966, pp. 296-305.

20/ Kastenbaum, Robert, "The Mental Life of Dying Geriatric Patients," Proceedings 7th International Congress of Gerontology, Vienna, 6, 153-159, 1966.

21/ Kogan, Nathan, and Florence Shelton, "Images of 'Old People' and 'People in General' in an Older Sample," Journal of Genetic Psychology, 100, 1962, 3-21.

22/ Kuhlen, op. cit.,

23/ Gurin, Gerald, Joseph Veroff, and Sheila Feld, Americans View Their Mental Health, New York: Basic Books, 1960

24/ Dohrenwend, Barbara S., and Bruce P Dohrenwend, "Sources of Refusals in Surveys," Public Opinion Quarterly, 32 (1968) 74-83.

Ferber, Robert, "Item Nonresponse in a Consumer Survey," Public Opinion Quarterly, 30 (1966), 399-415.

Mercer, Jane R., and Edgar W. Butler, "Disengagement of the Aged Population and Response Differentials in Survey Research," Social Forces, 46 (1967) 89-96.

Loewenstein, Regina, John Colombotos, and Jack Elinson, "Interviews Hardest to Obtain in an Urban Health Survey," Proceedings of the Social Statistics Section of the American Statistical Association. Washington: American Statistical Association, 1962, pp. 160-164.

25/ Kivlin, Joseph E., "Contributions to the Study of Mail-Back Bias," Rural Sociology, 30 (1965) 322-326.

26/ Dakin, Ralph E., and Donald Tennant, "Consistency of Response by Event-Recall Intervals and Characteristics of Respondents," The Sociological Quarterly, 9 (1968) 73-84.

27/ Gergen, Kenneth J., and Kurt W. Back, "Communication in the Interview and the Disengaged Respondent," Public Opinion Quarterly, 30 (1966) 385-398.

UNRELATED QUESTION RANDOMIZED RESPONSE TECHNIQUES WITH TWO TRIALS PER RESPONDENT[1/]

A. L. Gould and B. V. Shah, Research Triangle Institute
and
J. R. Abernathy, University of North Carolina

1. Introduction

Surveys aimed at obtaining information about the incidences of abortions in a population, the proportion of births occurring to unwed mothers, etc., usually are complicated by the fact that respondents are often reluctant to truthfully answer direct questions about such sensitive subjects. In an effort to circumvent this problem, and elicit a greater degree of cooperation from the respondents, researchers have developed randomized response techniques [1-4]. With such techniques, interviewed individuals are required to answer a randomly chosen question, with only the answer and not the question answered being known to the interviewer. However, even with randomized response procedures, respondents may refuse to cooperate. The purpose of the research reported here is to develop some models for the behavior of respondents when the unrelated question randomized response technique is used, and to apply these models to the results from surveys aimed at obtaining information about sensitive subjects; in particular we shall consider the results from a survey aimed at obtaining information about the proportion of households in North Carolina in which an illegitimate birth occurred [3].

Randomized response techniques originate from the work of Warner [4], who proposed that the sensitive question be presented to a respondent in a negative and a positive form, with the respondent randomly choosing one of the two forms and (truthfully) answering the chosen question without telling the interviewer which question was being answered. According to Warner, in view of the possibility of untruthful answers to direct questions, appreciable increases in efficiency are realizable with the randomized response technique.

The unrelated question randomized response variation of Warner's model was first reported by Horvitz, Shah, and Simmons [3], and its theoretical framework has been discussed by Greenberg, Abul-Ela, Simmons, and Horvitz [2]. The purpose of the variation was to increase the cooperation of respondents and the veracity of their responses beyond what the Warner technique might accomplish by, for example, removing the "heads I win, tails you lose" impression the structure of the Warner procedure might give. The unrelated question procedure poses two questions to the respondent, one dealing with the sensitive subject and the other being unrelated and innocuous; a randomizing device determines which question is to be answered, and the respondent (hopefully truthfully) answers the specified question without telling the interviewer which question is being answered. Horvitz, Shah and Simmons considered the situations where the respondent uses the device twice, giving two (assumed) independent answers, or once, giving one answer. They developed the formulas when one and two independent samples are interviewed, with the probability that the sensitive question will be presented to the respondent being different in the two samples, and applied their results to data from a study on the proportion of households in which an illegitimate birth occurred. The predicted proportions of households with illegitimate births differed from what would be expected, and they discussed some reasons for the discrepancies, observing that alternative models for the behavior of the randomizing devices and respondents might be appropriate. Such alternative models form the subject matter of this paper, and the data reported by Horvitz, Shah, and Simmons will be used subsequently for application of the various models to be discussed.

2. Preliminary Considerations

The unrelated question randomized response model assumes that the randomizing device presents the sensitive question to a respondent with the nominal probability, and that each respondent truthfully answers the presented question. In practice, these assumptions probably do not hold because of the idiosyncrasies of human beings and the imperfection of mechanical devices.

For example, respondents may not understand the procedure even after having it explained to them, and instead of answering the presented question, may randomly answer affirmatively or negatively. Alternatively, respondents in the sensitive group may, with a certain probability, answer the sensitive question untruthfully when it is presented, and this probability may depend (if each respondent takes two trials at using the device) on what had transpired previously. (That is, whether this was the first or second time the device was being used or the sensitive question was being presented, and if the second time, what had occurred on the first trial.) As still another possibility, a respondent in the sensitive group may refuse with some probability to truthfully answer any question requiring an affirmative answer, figuring perhaps that any positive answer might be stigmatizing. Or, possibly, this might be the attitude of respondents even in the non-sensitive group. Yet another possibility is that, when presented with the sensitive question, a respondent might decide to (possibly, but not necessarily, truthfully) answer the non-sensitive question.

Besides all the multitude of behavioral variations on the part of respondents, the randomizing devices may not be presenting the two questions with the nominal (assumed) probabilities. Horvitz, Shah, and Simmons discussed a number of possible explanations for this situation.

A general model from which particular behavioral models may be obtained by introducing sets of assumptions turns out to be a most practical vehicle for considering the

1/ Research supported by the National Institutes of Health, Contract No. 325-NIH-918A.

possible variations in behavior in a unified way. To specify such a model, some basic assumptions about the behavior of respondents are required; for the model to be considered, only four assumptions are required:

(1) For the two trial situation, a respondent who does (not) understand the procedure on the first trial will (not) understand the procedure on the second trial;

(2) Any respondent will always answer a question truthfully if the true answer is the negative (that is, the questions are defined so that a negative answer can never be stigmatizing);

(3) The randomizing device is such that the probability either question is selected at any trial does not depend upon which questions were selected at a previous trial (that is, separate trials with the randomizing device are independent and identically distributed).

(4) The innocuous question is unrelated to the sensitive question (i.e., the probability that an individual is in one or the other of the groups specified by the innocuous question does not depend upon the individual's possessing or not possessing the sensitive attribute, and vice versa).

To develop the probabilities of the various responses an individual may make, no other basic assumptions are required.

3. The General Model

We shall consider the situation in which each respondent uses the randomizing device twice. The form of the model which applies when there are two trials per respondent is given in Figure 1. The various probabilities appearing in Figure 1 are conditional upon events which precede them along the flow chart (which starts in the diamond labelled "comprehend?"). Aside from the parameter p which characterizes the behavior of randomizing device (and of which there are as many values as there are samples), the double trial model contains forty-two paramaters. Not all of these parameters are simultaneously estimable; indeed, given two samples in the two-trial case, at most six parameters can be estimated. However, by judiciously selecting the paramters to be estimated and the assumptions (fixed values, equality relations, etc.) to be applied to the remaining parameters, a wide variety of models for which the parameters can be estimated may be defined.

For an individual from whom two responses are elicited there are four possible responses: "yes, yes", "yes, no", "no, yes", "no, no". By following along the flow chart in Figure 1 the expressions for the various response probabilities may be obtained; these are given as Equations (1).

The parameter p appearing in these expressions denotes the probability with which the randomizing device presents the sensitive question to a respondent. If two samples from the population are drawn, different randomizing devices, with different values of p, will generally be used for the two samples. In what follows, p_1 will denote the value of p for the randomizing devices used in the first sample, and p_2 will denote the value of p for the randomizing devices used in the second sample. The parameter π_1 denotes the proportion of individuals in the population who possess the sensitive attribute, the parameter π_2 denotes the proportion possessing the innocuous attribute, and the parameter π_3 denotes the proportion who do not understand the procedure and answer at random, independently of the attributes they do or do not possess (with the ρ's denoting the probabilities of the various answers for

EQUATIONS (1)

$$\Pr\{\text{yes, yes}\} = \Big\{p^2\{\pi_1\xi_1\xi_2\sigma_1\sigma_2 + \pi_1\pi_2[\xi_1(1-\xi_2)\sigma_1\sigma_3 + (1-\xi_1)\xi_4\mu_1\mu_2 + (1-\xi_1)(1-\xi_4)\mu_1\mu_3]\} + p(1-p)\pi_1\pi_2\{\xi_1\sigma_1\sigma_4 + (1-\xi_1)\mu_1\mu_4 + \xi_7\nu_1\nu_2 + (1-\xi_7)\nu_1\nu_3\} + (1-p)^2\{\pi_1\pi_2\nu_1\nu_4 + (1-\pi_1)\pi_2\eta_1\eta_3\}\Big\}\pi_3 + \rho_1\rho_2(1-\pi_3)$$

$$\Pr\{\text{yes, no}\} = \Big\{p^2\{\pi_1\xi_1\xi_2\sigma_1(1-\sigma_2) + \pi_1\pi_2[\xi_1(1-\xi_2)\sigma_1(1-\sigma_3) + (1-\xi_1)\xi_4\mu_1(1-\mu_2) + (1-\xi_1)(1-\xi_4)\mu_1(1-\mu_3)] + \pi_1(1-\pi_2)\xi_1(1-\xi_2)\sigma_1\} + p(1-p)\{\pi_1\pi_2[\xi_1\sigma_1(1-\sigma_4) + (1-\xi_1)\mu_1(1-\mu_4) + \xi_7\nu_1(1-\nu_2) + (1-\xi_7)\nu_1(1-\nu_3)] + \pi_1(1-\pi_2)\xi_1\sigma_1 + (1-\pi_1)\pi_2\eta_1\} + (1-p)^2\{\pi_1\pi_2\nu_1(1-\nu_4) + (1-\pi_1)\pi_2\eta_1(1-\eta_3)\}\Big\}\pi_3 + \rho_1(1-\rho_2)(1-\pi_3)$$

$$\Pr\{\text{no, yes}\} = \Big\{p^2\{\pi_1(1-\sigma_1)\sigma_5\xi_1\xi_3 + \pi_1\pi_2[\xi_1(1-\xi_3)(1-\sigma_1)\sigma_6 + (1-\xi_1)\xi_5(1-\mu_1)\mu_5 + (1-\xi_1)(1-\xi_5)(1-\mu_1)\mu_6] + \pi_1(1-\pi_2)(1-\xi_1)\xi_6\mu_8\} + p(1-p)\{\pi_1\pi_2[\xi_1(1-\sigma_1)\sigma_7 + (1-\xi_1)(1-\mu_1)\mu_7 + \xi_8(1-\nu_1)\nu_5 + (1-\xi_8)(1-\nu_1)\nu_6] + \pi_1(1-\pi_2)\xi_9\nu_8 + (1-\pi_1)\pi_2\eta_2\} + (1-p)^2\{\pi_1\pi_2(1-\nu_1)\nu_7 + (1-\pi_1)\pi_2(1-\eta_1)\eta_4\}\Big\}\pi_3 + (1-\rho_1)\rho_3(1-\pi_3)$$

$$\Pr\{\text{no, no}\} = 1 - \Pr\{\text{yes, yes}\} - \Pr\{\text{yes, no}\} - \Pr\{\text{no, yes}\}$$

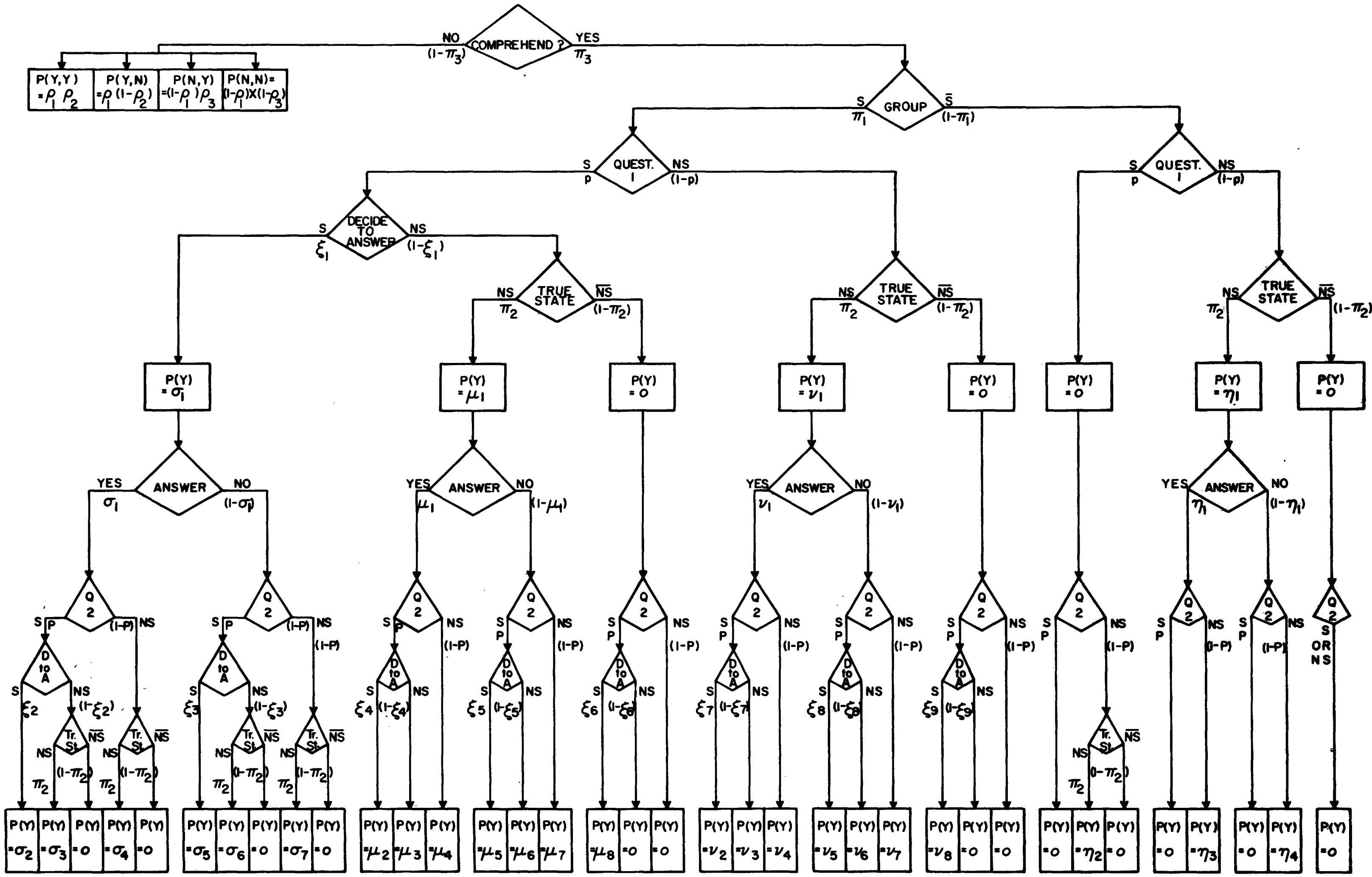


Figure 1: Flow Diagram of the Behavioral Model. (S means the sensitive question or group, NS means the nonsensitive question or group, and a bar over S or NS means the group in the population not possessing the attribute specified by the symbol.)

these individuals). The ξ parameters denote probabilities that an individual possessing the sensitive attribute will, when presented with the sensitive question, choose to consider instead the innocuous question. The remaining parameters denote the probabilities of telling the truth in various situations. If the responses of different individuals are independent, then the set of response probabilities (1) constitute the parameters of a multinomial distribution. Table 1 contains the probability functions, with the conbinational factors omitted, appropriate to the two-trial design for one and two samples.

The complicated functional forms of the response probabilities for the models render attempts to obtain explicit functional forms for estimators of the parameters impractical, if not impossible. Even for relatively simple behavioral models involving only a few of the parameters, explicit solutions for the parameters may not be unique. In order to make consideration of a variety of models convenient, the parameters were estimated by directly maximizing the likelihood function, using a direct search technique, with the parameters to be estimated and the various constraints specified in the input. Unfortunately, a direct maximization of the likelihood function does not lead to an expression of the covariance matrix of estimators as a by-product, as, for example, Gauss-Newton iteration does. Another approach to obtaining the covariance matrix of estimates is required; although results are not yet available, the technique of pseudoreplication will be used to obtain the estimates of the variances and covariances of the estimators. In view of the desirability of looking at the variation in values of the likelihood function [5], this technique has considerable merit. It proceeds by randomly allocating the units of samples into subsets, and estimating the parameters separately for each subset. The variances and covariances of the estimators obtained from the whole sample are then estimated using the corresponding quantities computed for the estimates obtained separately for each of the subsets.

4. Behavioral Models Obtained from the General Model.

The behavioral models to be applied to the illegitimacy data are described in Table 2. The general model is so formulated that, except for the p's, π's, ξ's, and ρ's, all the parameters are probabilities that a "yes" answer will be given when it is the truth; a "no" answer is assumed to be always given when it is the truth. In performing the calculations, all parameters which do not have specified fixed values, are being solved for, are not otherwise equal to other parameters, or do not satisfy other constraints, are assumed to have a value equal to unity.

TABLE 1

PROBABILITY FUNCTIONS OF THE POSSIBLE RESPONSES WHEN RESPONSES OF DIFFERENT INDIVIDUALS ARE ASSUMED INDEPENDENT; TWO TRIALS PER RESPONDENT.*)

Structure of Survey	Notation	Probability Function, Combinatorial Factors Omitted
Two trials per respondent, one sample	n_{11} = no. of "yes, yes" responses n_{12} = no. of "yes, no" responses n_{01} = no. of "no, yes" responses	$\alpha^{n_{11}}\beta^{n_{10}}\gamma^{n_{01}}(1-\alpha-\beta-\gamma)^{n-n_{11}-n_{10}-n_{01}}$
Two trials per respondent, two samples	n_{11} = no. of "yes, yes" responses in sample 1 n_{10} = no. of "yes, no" responses in sample 1 m_{11}, m_{10}, m_{01} are the corresponding quantities for sample 2	$\alpha_1^{n_{11}}\beta_1^{n_{10}}\gamma_1^{n_{01}}$ $\cdot\ (1-\alpha_1-\beta_1-\gamma_1)^{n-n_{11}-n_{10}-n_{01}}$ $\cdot\ \alpha_2^{m_{11}}\beta_2^{m_{10}}\gamma_2^{m_{01}}$ $\cdot\ (1-\alpha_2-\beta_2-\gamma_2)^{m-m_{11}-m_{10}-m_{01}}$

*) α = Pr{yes, yes}, β = Pr{yes, no}, γ = Pr{no, yes}; the subscripts of α, β, γ denote the sample to which the values of the probabilities apply.

TABLE 2

SOME BEHAVIORAL MODELS WHICH MAY BE OBTAINED FROM THE GENERAL MODEL

Model Number	Verbal Description	Parameters to be Estimated	Fixed Values Not Equal to Unity	Equalities	Other Constraints
1	All respondents comprehend, and answer the presented question truthfully	π_1, π_2	$p_1 = .7$ $p_2 = .3$	$\rho_2 = \rho_1$ $\rho_3 = \rho_1$	
2	Not all respondents comprehend; those who do answer the presented question truthfully; those who don't are equally likely to answer "yes" or "no"	π_1, π_2, π_3	$p_1 = .7$ $p_2 = .3$ $\rho_1 = .5$		
3	Not all respondents comprehend; those who do answer the presented question truthfully; those who don't answer "yes" with unknown probability	π_1, π_2, π_3 ρ_1	$p_1 = .7$ $p_2 = .3$	$\rho_2 = \rho_1$ $\rho_3 = \rho_1$	
4	Same as for Model 1	π_1, π_2, p_1			$p_2 = .42857\ p_1$
5	Same as for Model 1	π_1, π_2, p_1			$p_2 = p_1 - .4$
6	Same as for Model 1	π_1, π_2, p_1, p_2			
7	All respondents comprehend, respondents in the sensitive group might lie when presented with the sensitive question, and this doesn't depend upon any previous behavior; otherwise, the presented question is answered truthfully	π_1, π_2, σ_1	$p_1 = .7$ $p_2 = .3$	$\sigma_2 = \sigma_1$ $\sigma_5 = \sigma_1$ $\nu_2 = \sigma_1$ $\nu_8 = \sigma_1$	
8	All respondents comprehend; respondents in the sensitive group might lie (with unknown probability) when presented with the sensitive question the first time; if presented with the sensitive question a second time, they lie with probability 1 if they lied the first time, and with unknown probability (possibly different from the first) if they told the truth the first time; otherwise, the presented question is answered truthfully)	π_1, π_2, σ_1, σ_2	$\sigma_5 = 0$ $p_1 = .7$ $p_2 = .3$	$\nu_2 = \sigma_1$ $\nu_8 = \sigma_1$	
9	Same as for 8, except that respondents in the sensitive group may answer the sensitive question truthfully if it is presented a second time and they lied the first time it was presented	π_1, π_2, σ_1, σ_2, σ_5	$p_1 = .7$ $p_2 = .3$	$\nu_2 = \sigma_1$ $\nu_8 = \sigma_1$	

Table 2 (cont'd)

Model Number	Verbal Description	Parameters to be Estimated	Fixed Values Not Equal to Unity	Equalities	Other Constraints
10	All respondents comprehend; when respondents in the sensitive group are presented with the sensitive question, they might lie; when respondents are required to answer the nonsensitive question affirmatively, they might lie, possibly with a different probability	$\pi_1, \pi_2, \sigma_1, \sigma_4$	$p_1 = .7$ $p_2 = .3$	$\sigma_2 = \sigma_1$ $\sigma_5 = \sigma_1$ $\nu_2 = \sigma_1$ $\nu_5 = \sigma_1$ $\nu_1 = \sigma_4$ $\nu_4 = \sigma_4$ $\nu_7 = \sigma_4$ $\eta_1 = \sigma_4$ $\eta_2 = \sigma_4$ $\eta_3 = \sigma_4$ $\eta_4 = \sigma_4$	
11	All respondents comprehend; any respondent faced with giving an affirmative answer to a presented question might lie	π_1, π_2, σ_1	$p_1 = .7$ $p_2 = .3$	$\sigma_2=\sigma_4=\sigma_5=\sigma_7$ $=\nu_1=\nu_2=\nu_4=\nu_5$ $=\nu_7=\eta_1=\eta_2=\eta_3$ $=\eta_4=\nu_8=\sigma_1$	
12	All respondents comprehend; respondents in the sensitive group, when presented with the sensitive question, might decide instead to truthfully answer the nonsensitive question; the nonsensitive question is always truthfully answered	π_1, π_2, ξ_1	$p_1 = .7$ $p_2 = .3$	$\xi_2 = \xi_3 = \xi_4$ $= \xi_5 = \xi_6$ $= \xi_7 = \xi_8$ $= \xi_9 = \xi_1$	

5. Application of the Models to Data from a Study on the Proportion of Households in North Carolina in which an Illegitimate Birth Occurred.

A summary of the data from the illegitimacy study reported by Horvitz, Shah, and Simmons [3], is presented in Table 3.

For all of the surveys, p_1 equals 0.7 and $p_2 = 0.3$. A summary of the results of the calculations is presented in Table 4. Table 5 contains a summary of the results of Horvitz, Shah, and Simmons.

TABLE 3

FREQUENCIES OF THE VARIOUS RESPONSES FROM THE ILLEGITIMACY STUDY. THE OBJECTIVE OF THE STUDY WAS TO DETERMINE THE PROPORTION OF HOUSEHOLDS IN NORTH CAROLINA REPORTING A BIRTH TO AN UNWED MOTHER. TWO RANDOMIZING DEVICES WERE USED. SEE [3] FOR DETAILS.

POPULATION SUBSET	Sample 1, p = .7				Sample 2, p = .3			
	n_{11}	n_{10}	n_{01}	n	m_{11}	m_{10}	m_{01}	m
White households (randomizing device is deck of cards)	137	271	253	1227	512	291	215	1340
Nonwhite households (same randomizing device)	29	52	45	223	124	54	61	298
White households (randomizing device is bead box)	37	55	61	320	141	67	48	375
Nonwhite households (same randomizing device)	16	21	22	117	25	13	5	67

TABLE 4.

SUMMARY OF CALCULATIONS FOR THE ILLEGITIMACY DATA FROM [3]*

Model Number (See TABLE 3)

		1	2	3	4	5	6	7	8	9	10	11	12
No. Parameters		2	3	4	3	3	4	3	4	5	4	3	3
White HH. Card Deck	Log Likelihood	-3368.42	-3346.40	-3345.83	-3357.82	-3367.56	-3344.47	-3352.64	-3362.34	-3351.51	-3348.65	-3364.36	-3368.42
	Chi-Square	64.88	19.28	17.92	43.96	63.23	15.63	33.82	52.92	31.65	24.17	56.51	64.88
	d.f. (χ^2)	4	3	2	3	3	2	3	2	1	2	3	3
	$\hat{\pi}_1$	.02824	.004041	.01872	.00211	.02005	.002749	.9482	.08408	.2181	.3931	.05869	.02865
	$\hat{\pi}_2$	.8616	.8514	.8925	.8391	.8531	.8637	.8134	.8427	.8074	.8296	.8854	.8618
Non-White HH. Card Deck	Log Likelihood	-685.84	-682.02	-683.01	-683.94	-685.76	-681.53	-685.72	-684.87	-682.88	-681.55	-685.28	-685.84
	Chi-Square	10.00	2.23	2.20	6.16	9.79	1.24	9.76	8.00	4.00	1.30	8.97	10.00
	d.f. (χ^2)	4	3	2	3	3	2	3	2	1	2	3	3
	$\hat{\pi}_1$	.04299	.02163	.01989	.01995	.02053	.01057	.04805	.09615	.2322	.5264	.05727	.04299
	$\hat{\pi}_2$	.8981	.9073	.9078	.8781	.8877	.8979	.8958	.8837	.8540	.8553	.9088	.8978
White HH. Bead Box	Log Likelihood	-879.56	-878.16	-877.90	-877.61	-879.29	-876.88	-879.46	-879.52	-879.45	-879.51	-879.54	-879.55
	Chi-Square	9.33	6.06	5.66	5.48	8.89	3.81	9.12	9.25	9.11	9.14	9.20	9.30
	d.f. (χ^2)	4	3	2	3	3	2	3	2	1	2	3	3
	$\hat{\pi}_1$	.05434	.01999	.02052	.02024	.02024	.02038	.05830	.05767	.05943	.06752	.06533	.05446
	$\hat{\pi}_2$	.7646	.7478	.7326	.7446	.7518	.7502	.7631	.7631	.7624	.7688	.7702	.7646
Non-White HH. Bead Box	Log Likelihood	-232.76	-231.76	-231.76	-231.09	-231.99	-230.93	-230.80	-232.07	-230.79	-231.31	-232.77	-232.75
	Chi-Square	7.28	4.78	4.79	4.32	6.22	3.70	3.55	6.25	3.54	4.31	7.26	7.27
	d.f. (χ^2)	4	3	2	3	3	2	3	2	1	2	3	3
	$\hat{\pi}_1$	.08512	.06436	.02098	.004756	.04941	.004050	.5374	.1594	.3059	.3305	.08888	.08596
	$\hat{\pi}_2$	.7525	.7730	.6758	.7048	.7188	.7079	.6464	.7120	.6438	.6639	.7532	.7517

* The sensitive attribute is occurrence of an illigitimate birth in the household during the past year; the nonsensitive attribute is having been born in North Carolina.

TABLE 5

SUMMARY OF RESULTS OF HORVITZ, SHAH AND SIMMONS

Population	Expected $\hat{\pi}_1$	Model 1 Estimates $\hat{\pi}_1$	Model 1 Estimates $\hat{\pi}_2$	Model 4 Estimate $\hat{\pi}_1$
White HH. Card Deck	.002	.141	.755	-.0003
Nonwhite HH. Card Deck	.034	.151	.805	.027
White HH. Bead Box	.002	.122	.704	-.0007
Nonwhite HH. Bead Box	.034	.180	.648	.020

6. Discussion and Remarks

The estimates obtained by Horvitz, Shah and Simmons differ from those obtained through the use of the model because the Horvitz, Shah and Simmons estimators are moment estimators, and those obtained for the model are obtained by a search for a set of values maximizing the likelihood function. In discussing the large deviation of the estimate of π_1 from its expected value, Horvitz, Shah and Simmons considered a number of explanations relevant to the modeled behavior, in particular the possibility that the realized p values were different from the nominal ones, and that there may have been some confusion on the part of the respondents. They considered also the possibility that misreading of the sensitive question, or the way the procedure was described to the respondents, may have led to an excess of positive answers. The possibility in the general model which comes closest to describing this source of error is model 3, although the situation is not exactly that of confusion on the part of the respondents. Deliberate untruthfulness on the part of the respondents does not appear to be a relevant explanation in the present case. From the results of fitting the models, it appears that models 3 and 6 do the best job in describing the data. The estimated parameter values and predicted frequencies for these models are displayed in Table 6. We note that the predicted proportions of households with an illegitimate birth are considerably closer to what would be expected than was the case for Horvitz, Shah

TABLE 6

PARAMETER ESTIMATES AND PREDICTED

FREQUENCIES FOR MODELS 3 AND 6

Population	MODEL 3 $\hat{\pi}_1$	MODEL 3 $\hat{\pi}_2$	MODEL 3 $\hat{\pi}_3$	MODEL 3 $\hat{\rho}_1$	MODEL 6 $\hat{\pi}_1$	MODEL 6 $\hat{\pi}_2$	MODEL 6 $\hat{p}_1$	MODEL 6 $\hat{p}_2$
White HH. Card Deck	.01872	.8925	.7179	.4102	.002749	.8637	.6270	.3373
Nonwhite HH. Card Deck	.01989	.9078	.7462	.5000	.01057	.8979	.6208	.3241
White HH. Bead Box	.02052	.7326	.8448	.5780	.02038	.7502	.6222	.2995
Nonwhite HH. Bead Box	.02098	.6758	.7509	.5898	.004050	.7090	.5548	.2719

Population	Sample	Observed YY	Observed YN	Observed NY	Observed NN	Model 3 YY	Model 3 YN	Model 3 NY	Model 3 NN	Model 6 YY	Model 6 YN	Model 6 NY	Model 6 NN
White HH. Card Deck	1	137	271	253	566	143	246	246	592	150	247	247	582
	2	512	291	215	322	493	269	269	310	510	258	258	314
Nonwhite HH. Card Deck	1	29	52	45	97	31	45	45	102	31	47	47	99
	2	124	54	61	59	120	61	61	57	124	58	58	58
White HH. Bead Box	1	37	55	61	167	39	53	53	175	39	56	56	170
	2	141	67	48	119	136	62	62	117	141	58	58	117
Nonwhite HH. Bead Box	1	16	21	22	58	17	19	19	61	17	20	20	59
	2	25	13	5	25	23	11	11	22	25	9	9	23

and Simmons' results; this may be due to the fact that the present estimators are maximum likelihood, rather then moment estimators. The relatively close agreement of the predicted frequencies with the observed frequencies does not, of course, mean that the models are adequate representations of reality: models 3 and 6 describe different types of behavior. However, the models are relatively consistent insofar as their predictions of the values of π_1 are concerned. At least for the population considered, the percentage of households in which an illegitimate birth occurred appears to be approximately 2%, for both white and non-white households. However, the precisions of the estimates remain to be determined.

There is certainly a need for considerable empirical experience in applying the models to surveys employing the unrelated question randomized response technique. Such experience is necessary for determining which of the particular models obtainable from the general model will be useful in various situations, and which of the models will prove to be of limited or negligible utility. Moreover, empirical experience is also necessary to determine if there are behavioral models not obtainable from the general model which are useful for describing the behavior of respondents.

Acknowledgement

It is a pleasure to acknowledge the illuminating comments and discussion of Dr. D. G. Horvitz in bringing the problem to our attention, and during the course of carrying out the research.

7. References

[1] Abul-Ela, A. A., Greenberg, B. G., and Horvitz, D. G. (1967). A multi-proportions randomized response model. Jour. Am. Stat. Assoc. 62, 999-1008.

[2] Greenberg, B. G., Abul-Ela, A. A., Simmons, W. R., and Horvitz, D. G. (1969). The unrelated question randomized response model: theoretical framework. Jour. Am. Stat. Assoc. 64, 520-539.

[3] Horvitz, D. G., Shah, B. V., and Simmons, W. R. (1967). The unrelated question randomized response model. Social Statistics Section of the American Statistical Association, 65-72.

[4] Warner, S. L. (1965). Randomized response: a survey technique for eliminating evasive answer bias. Jour. Am. Stat. Assoc. 60, 63-69.

[5] Sprott, D. A., and Kalbfleisch, J. D. (1969). Examples of likelihoods and comparison with point estimates and large sample approximations. Jour. Am. Stat. Assoc. 64, 468-484.

A GENERAL RESPONSE AND MEASUREMENT ERROR MODEL AND ITS APPLICATION TO THE ANALYSIS OF 2 x 2 CONTINGENCY TABLES[1/]

A. Lawrence Gould, Research Triangle Institute

1. Introduction

Nonsampling errors in survey data have been the subject of considerable interest and research, as the extensive bibliography in Cochran [1] indicates. Errors of this sort may be classed as nonresponse errors, where for one reason or another observed measurements on some of the sample units are not available, and what are often called "measurement" errors; this paper is concerned with these latter errors.

Treatments of "measurement" errors tend to take one of two approaches: a model is postulated for their probability structure and the consequences with regard to the distributions of various statistics are worked out; or, starting with reinterview data, various statistics are devised to indicate the effects of the "measurement" errors, with little explicit specification of a probability structure for the errors. The research discussed is an attempt to tie these approaches together by developing a general model for response and measurement errors, and showing how the elements of this model enter into the distribution of statistics developed through the second approach just mentioned.

2. The General Model

Suppose there is a population of N discrete individuals, and that with each individual there is associated a vector, ξ , of characteristics. Usually, for an individual, the elements of ξ are regarded as being fixed; however, there are circumstances in which it is more appropriate to regard them as random variables. If the elements of ξ are regarded as fixed, the measurement errors have the interpretation that on repeated attempts to observe the value of ξ , different realized values would occur. The elements of ξ may be regarded appropriately as random when, for instance, they correspond to states of mind of an individual with regard to various political or social issues. When a survey is conducted at some time (or trial) t , and a sampled individual is approached for interview or measurement, ξ assumes a particular value ξ_t . The distribution of potential values of ξ may be different for different individuals, although they will be assumed to be in the same family of distributions, so that the individual distributions will differ only in the values of the elements of a parameter vector, θ . The distribution (probability or density function) of ξ over all trials t , for the i-th population member will be denoted by $f(\xi;\theta_i)$. In turn, θ may be regarded as having a distribution over different individuals of the population, with probability or density function $g(\theta;\boldsymbol{\Theta})$, $\boldsymbol{\Theta}$ being a set of parameters characteristic of the population. The act of drawing a sample of individuals from the population is therefore equivalent to drawing a sample of θ's, and also equivalent to drawing a sample of ξ distributions.

1/ Research supported by the U. S. Bureau of the Census, Contract No. Cco-9319.

When a survey is conducted under conditions γ at trial t , the i-th member of the population (if in the sample) will yield a measured response η_{it} when information regarding ξ is elicited. If the measuring or interviewing process contains no errors, then $\eta_{it} = \xi_{it}$, the value of ξ at trial t for the i-th individual. This is, however, only the ideal situation; in practice, η_{it} may be regarded as a random variable with a distribution depending upon ξ_{it} and the conditions under which the survey is conducted. The probability or density function of η_{it} (the evoked response of individual i at trial t may be denoted by

$$h_\gamma(\gamma;\ \xi_{it},\ \phi_i(\gamma)) \quad ,$$

where $\phi_i(\gamma)$ denotes a set of parameters specific to the i-th individual and survey conditions γ . As do the θ's , the ϕ's have a distribution over the population, with probability or density function $k_\gamma(\phi, \boldsymbol{\Phi}(\gamma))$. The notion of a "trial" t may be generalized to include a set of "trials" $\{t_1, \ldots, t_m\}$, in which case ξ_{it} and η_{it} are vectors of possibly correlated elements:

$$\xi_{it} = (\xi_{it_1}, \ldots, \xi_{it_m})'$$

$$\eta_{it} = (\eta_{it_1}, \ldots, \eta_{it_m})' .$$

Thus, repeated surveys may readily be dealt with in the framework of the model.

The distribution $f(\xi;\ \theta_i)$ may be thought of as the response error distribution for an individual because it describes the distribution of potential responses (ξ's) the individual might present at an interview. The distribution $h_\gamma(\eta;\ \xi_{it},\ \phi_i(\gamma))$ may be regarded as the (conditional) measurement error distribution which depends upon the survey conditions, the individual, and the response he presents at the interview; it describes the distribution of potential measurements of the response. For this reason, ξ will be called the "response", and η , the "measurement".

As a physical example to illustrate the distinction between these two conceptual errors, suppose one wishes to measure the length of a metal bar on two occasions, and that the unit of length is taken to be the meter bar formerly used as a length standard as it existed at a given instant of time. Then, at either occasion, the metal bar of interest has, in terms of this length standard, a definite length. The objective of the measurement procedure is to determine what the lengths on the two occasions are. Because of environmental differences between the two occasions of measurement, the "true" length of the bar may not be the same on both occasions; this variation is the response error. On measuring the bar's length on either occasion, the inaccuracy of the measuring instruments may lead to an observed length different from the "true" length at that occasion; this variation is the "measurement" error.

As well as the distribution of measurements on an individual conditional upon a particular response ξ_t, one may envisage a distribution for the measurements averaged over all possible responses under a given set of survey conditions. Such a distribution is

$$h^*_\gamma(\eta;\ \theta_i,\ \phi_i(\gamma)) = \begin{cases} \int_\xi h_\gamma(\eta;\ \xi,\ \phi_i(\gamma))\ f(\xi;\ \theta_i)\ d\xi \\ \\ \sum_\xi h_\gamma(\eta;\ \xi,\ \phi_i(\gamma))\ f(\xi;\ \theta_i) \end{cases},$$

depending upon whether ξ is a continuous or discrete random variable. Intuitively, h^* may be regarded as follows: given a particular individual, if one could obtain a measurement η at various trials under the same survey conditions and the act of measuring did not affect the process generating the measurements, then the observed distribution of measurements would be described by h^*_γ.

The distributions f, h, and h^* are marginal distributions. In reality it may happen that the responses or measurements of different individuals are correlated; this is essentially the point at which the two approaches to analyzing non-sampling errors of the "measurement" type diverge. Generally speaking, population structure approaches assume the between-individual correlations to be either absent or to have a particular form; the sample statistic-oriented approach makes no assumption about the absence of correlations between individuals.

The remainder of this paper is concerned with an application of the model to the analysis of 2 x 2 contingency tables. In this situation, an individual is in, or is assigned to, one of four classes according to his possession or non-possession of either of two attributes. For notational convenience, an individual's response and measurement may be regarded as 4-element vectors; the j-th element of a vector being unity if and only if the individual is in, or is assigned to, the j-th of the four possible cells of the contingency table, and the remaining elements of the vector being zero.

3. An Example of the Population Structure Approach

From the population structure point of view, the roles of the response and measurement error distributions are so defined as to yield, for one or more trials, a presumed distribution for the observed cell frequencies. In application, then, the problem becomes that of estimating the parameters appearing in the functional form of the presumed distribution.

As a simple example of a population structure model for the 2 x 2 contingency table situation, assume that observed classifications of different individuals are independent, as are the classifications of any individual on separate trials. Then

$$f(\xi;\ \theta_i) = \xi'\theta_i\ ,$$

$$h(\eta;\ \xi_i,\ \phi_i) = \eta'\phi_i\xi_i\ ,$$

$$h^*(\eta;\ \theta_i,\ \phi_i) = \eta'\phi_i\theta_i$$

(ϕ is a matrix; all the other entities are vectors).

Suppose θ_i and ϕ_i are independently distributed, and that $\theta^* = E(\theta)$ and $\phi^* = E(\phi)$; then the averages of the frequency distributions f and h^* over all individuals in the population are

$$f'(\xi;\ \theta^*) = \xi'\theta^*$$

and

$$h^{*\prime}(\eta;\ \theta,^*\ \phi^*) = \eta'\phi^*\theta^*\ .$$

On the average basis, $h^{*\prime}$ may be taken as the distribution of the various cells in the table, and the distribution of observed frequencies is a multinomial distribution with the four possible values of $h^{*\prime}$ as the parameters. The objective is, given a set of observed classifications, to estimate the values of the elements of ϕ^* and θ^*. If the sample units are classified on two occasions, and the responses of the individuals are assumed not to change between the two trials, then the average joint distribution of the cells is

$$h^{*\prime}(\eta_1,\ \eta_2;\theta^*,\ \phi^*) = \eta_1'\ \phi^* \begin{bmatrix} \theta_1^* & & & 0 \\ & \theta_2^* & & \\ & & \theta_3^* & \\ 0 & & & \theta_4^* \end{bmatrix} \phi^{*\prime}\eta_2\ .$$

The models developed by Giesbrecht [3] and by Koch [6] lead to distributions which are special cases of (1) in that additional assumptions regarding the elements of ϕ^* are imposed. The population structure model developed in the example used could be generalized by altering the independence assumptions.

4. An Example of the Sample-Oriented Approach

The sample-oriented approach is by and large concerned with estimating the population proportions falling into the various categories, and with investigating the roles of various sources of error and various intercorrelations on the precision of the estimators of these proportions. Suppose that the elements of Θ represent the true proportions of the population falling into the various cells of the table, where Θ is defined by

$$\Theta = E\{\xi\} = \sum_\theta \sum_\xi \xi\ f(\xi;\ \theta)\ g(\theta;\Theta)\ .$$

In the presence of measurement errors, an unbiased estimate of Θ may not be obtainable. The expected value of a single observation on individual i of the population is

$$E\{\eta_i\} = \boldsymbol{\theta} + \sum_\theta \sum_\xi \left(\sum_\eta \eta\, h_\gamma(\eta_i; \xi, \phi_i(\gamma)) - \xi\right) \cdot f(\xi;\, \theta)\, g(\theta;\, \boldsymbol{\theta}) \;;$$

in general, the value of the bias term is not known. Given a simple random sample of n units from the population, the usual estimator of $\boldsymbol{\theta}$ is $\bar{Y}$, the sample mean. Other sampling schemes could be employed as well; the computations with simple random sampling are probably the simplest. The covariance matrix of the elements of $\bar{Y}$ may be shown to have the following form, assuming the θ's and ϕ's to be independently distributed in the population:

$$V\{\bar{Y}\} = \frac{n+(N-1)a}{nN} \cdot \frac{1}{N} \sum_{i=1}^{N} \{\phi^*\theta_i\, \theta_i' - \text{diag}\,[(\phi^*\theta_i)_1, \ldots, (\phi^*\theta_i)_4]\} + \frac{a}{nN} \sum_{i=1}^{N} \phi^*(\theta_i - \theta^*)(\theta_i - \theta^*)'\phi^{*\prime} + \frac{n-a}{n}\left(1 - \frac{1}{N}\right) \cdot \frac{1}{N(N-1)} \cdot \sum_{i \neq i'} \sum E\{(\eta_i - \phi^*\theta_i)(\eta_{i'} - \phi^*\theta_{i'})'\} \qquad (2)$$

where $\theta^* = E\{\phi\}$, $\phi^* = E\{\phi\}$, and the distribution of η_i is h^*_γ; if the sampling is without replacement, $a = (N-n)/(N-1)$; if the sampling is with replacement, $a = 1$; in both cases, the sampling is simple random. The quantities appearing in (2) correspond to those defined in the example in the previous section. The first term of (2) arises from the combined effects of response and measurement errors; the second term reflects the sampling error, and the third term expresses the degree of correlation between observed classifications of individuals -- it is proportional to what is usually called the "within-trial correlation between different individuals," -- or at least would be if the quantities of interest were scalars rather than vectors.

In a practical situation, the objective is often to estimate the sampling and response-measurement contributions. For the present case, it may be shown that, if $N \ll n$ and the third term of (2) is zero (which is not, by the way, to say that the measurement of an individual is uncorrelated with that of an other individual), then the usual estimator of the covariance matrix of the elements of the sample mean is an approximately unbiased estimator of $V\{\bar{Y}\}$. Further, if a repetition of the survey is conducted under conditions identical to those of the original survey, and if measurements on different trials are assumed uncorrelated, then the quantity g,

$$g = \sum_{j=1}^{n} (Y_{jt} - Y_{jt'})(Y_{jt} - Y_{jt'})' \;,$$

where Y_{jt} is the observed measurement on the j-th sample unit, with expectation

$$E\{g\} = \frac{n}{N} \sum_{i=1}^{N} \left\{\text{diag}[(\phi^*\theta_i)_1, \ldots, (\phi^*\theta_i)_4] - \phi^*\theta_i\theta_i'\phi^{*\prime}\right\} \;,$$

provides the basis for an unbiased estimator of the response-measurement component of (2), namely $g/2n^2$. Thus, the moments of the quantities considered in the sample-oriented approach may be related to the distribution appearing in the general model.

The example discussed in this section is based upon Koch's [7] extension to the 2 x 2 contingency table case of the work of Hansen, Hurwitz, and Pritzker [5] which, in turn, is based on the model of Hansen, Hurwitz, and Bershad [4]. Felligi [2] discusses the application of the Hansen, Hurwitz, and Bershad model to interpenetrating sample and re-interview surveys, considering in detail the various intercorrelations which might arise.

5. Remarks

One purpose of the proposed model is to provide a conceptual way of considering non-sampling errors of the "measurement" type which allows the population-structure and sample-oriented approaches to considering these errors to be related to each other, with an explicit indication of the respects in which the two approaches differ. For the 2 x 2 contingency table case, this relating of the two approaches is fairly straightforward under some simplifying assumptions, as the preceding discussion indicates.

The examples presented by no means exhaust the possibilities for developing the general model, and a number of areas of extension may readily be visualized, for instance: considering the applicability of the model to a number of other measurement error models, including those discussed by Cochran and by Mandel [8]; extension of the general model to include the possibility of sampling plans more general than simple random sampling, or surveys in which only a subset of the sample is reinterviewed, or to include the effect of non-response, etc.

6. References

[1] Cochran, W. G. (1968). Errors of measurement in statistics. Technometrics 40, 637-666.

[2] Felligi, I. P. (1964). Response variance and its estimation. Jour. Am. Stat. Assoc. 59, 1016-1041.

[3] Giesbrecht, F. G. (1967). Evaluation of the effects of classification errors on measures of association in contingency tables. RTI Report SU-215, prepared for Bureau of Census, Contract Cco-9191.

[4] Hansen, M. H., Hurwitz, W. N., and Bershad, M. A. (1961). Measurement errors in cen-

suses and surveys. Bull. Int. Stat. Inst. 38, No. 2, 359-374.

[5] Hansen, M. H., Hurwitz, W. N., and Pritzker, L. (1964). The estimation and interpretation of gross differences and the simple response variance. Contributions to Statistics Presented to Professor P. C. Mahalanobis, Pergamon Press, Oxford.

[6] Koch, G. G. (1968). A simple model for misclassification errors in 2 x 2 contingency tables. RTI report SU-363, prepared for Bureau of Census, Contract Cco-9260.

[7] Koch, G. G. (1968). The effect of nonsampling errors on measures of association in 2 x 2 contingency tables. RTI Report SU-363, prepared for Bureau of Census, Contract Cco-9260.

[8] Mandel, J. (1959). The measuring process. Technometrics 1, 251-267.

METHODOLOGY STUDY FOR DETERMINING THE OPTIMUM RECALL PERIOD FOR THE REPORTING OF MOTOR VEHICLE ACCIDENTAL INJURIES

William S. Cash and Abigail J. Moss
National Center for Health Statistics

I. Introduction

Since the beginning of the National Health Interview Survey in 1957, annual estimates on the incidence of injuries resulting from all types of accidents have been obtained. The most recent injury data available are for the period July 1966-June 1967; the data indicate that during this period, an estimated 51.8 million persons, 26.9 per 100 persons in the civilian, noninstitutional population, were injured. Of this number, 3.5 million or 1.8 per 100 persons were injured in moving motor vehicle accidents. Even though the number of persons injured in moving motor vehicle accidents constitutes only 6.8 percent of the total injured population, these data have been of particular interest to data consumers. One reason for this interest is that motor vehicle injures are often of a more serious nature than other types of injures. For example, the proportion of motor vehicle injuries resulting in activity restriction and bed disability is markedly higher than for other types of injuries.[1]

During the past two years, the National Center for Health Statistics has experienced an increased demand for more reliable and detailed statistics on motor vehicle injuries and other factors relating to motor vehicle accidents. Part of this demand was the result of a greater public awareness of the paucity of motor vehicle safety standards employed at that time and the high number of traffic fatalities and personal injuries resulting from motor vehicle accidents. Since the Center had not previously collected motor vehicle injury data in sufficient detail to satisfy the requests being made, a decision was made to obtain more detailed information on motor vehicle injuries on the 1968 Health Interview Survey questionnaire. Prior to the actual data collection phase however, it appeared that an evaluation study should be conducted for the purpose of establishing new estimating procedures for motor vehicle data. This was thought to be necessary since the collection and sampling procedures used earlier to estimate the annual incidence of injuries would result in an exceedingly high sampling error, if used to derive annual estimates for more detailed motor vehicle data. This report describes the methodological aspects of this special study which took place between February and May 1967 and presents the findings which were later incorporated by the Health Interview Survey.

In the past, estimates on the incidence of all types of injuries have been obtained by collecting data among sample persons on only those injuries that occurred during the two-week period preceding the interview and then inflating the frequencies to obtain annual estimates. The collection of injury data has been limited to a two-week recall period primarily because some specific kinds of injuries a person may receive have such little impact that respondents may forget to report them if much time has elapsed between the date of the accident and the interview. The degree of impact an injury has on an individual could be expected to vary depending upon: (1) the severity of the injury involved, and (2) the circumstances of the accident which caused the injury or, more specifically, the type of accident that occurred. Therefore, if one were to hypothesize that injury-producing motor vehicle accidents have a greater impact on an individual than injuries obtained from some other kinds of accidents, it could be assumed that a respondent would be able to remember this type of injury for a longer time period and report it in a household interview even when a recall period longer than two weeks was used. By increasing the length of the recall period for motor vehicle injuries, the number of injuries reported would be increased. This, in turn, would have the effect of decreasing the sampling error, making it feasible to collect and publish motor vehicle injury data in greater detail.

With these considerations in mind, this evaluation study was specifically designed to answer the following two questions: (1) can the recall period for injuries resulting from motor vehicle accidents be increased without greatly affecting the respondent's ability to report such occurrences, and if this is possible, (2) what is the optimum length of recall for the reporting of motor vehicle injuries?

II. The Study Design

Description of the Survey Procedure. After considering the various alternatives, it was determined that the best method for evaluating the optimum recall period for the reporting of motor vehicle injuries was a record check study. Briefly, the Motor Vehicle Evaluation Study would consist of interviewing a sample of persons known to have been in an injury-producing motor vehicle accident at some time during the twelve-month period preceding the interview. Accident information obtained from the respondent at the time of the interview would then be compared with comparable data recorded on an official accident report form. Final analysis of the data would consist of a comparison between a person's injury status as recorded on the accident record and on the questionnaire used in the interview, as well as a comparison of the reporting of other details relating to the accident. Of primary interest for evaluation however, would be the relationship between the ability of the respondent to report motor vehicle injuries and the length of time between the motor vehicle accident and the interview.

The Sample. The Motor Vehicle Evaluation Study was conducted in the Research Triangle Area of North Carolina where the Division of Health Interview Statistics has an experimental field interview unit established for the purpose of conducting methodological studies of this kind. The sample design used for this study had the following features. First, the North Carolina Department of Motor Vehicles provided the Health Interview Survey staff with accident record punch cards for those accidents occurring in Durham, Orange, and Wake counties in that State during February 1966 to February 1967 which met all of the specifications listed below:

1. one or more of the persons involved in the accident were residents of Durham, Orange, or Wake counties

2. one or more persons in the accident were injured

3. one or more persons survived the accident

These punch cards were then divided into three strata according to the time interval between the date of the accident and the interview: Stratum I, less than 3 months; Stratum II, 3-6 months; and Stratum III, 7-12 months. Within these strata, the cards were sorted by:

1. whether or not a legal violation was involved, and

2. the most severe injury sustained in the accident as reported by the police officer who completed the official accident report form.

The sample for the Motor Vehicle Evaluation Study was then drawn by using a simple random (systematic) sample by strata. The sampling fraction for Strata I and III was 1/6 and for Stratum II was 1/5. In order to detect differences at the .05 significance level, it was estimated that approximately 500 accidents would need to be selected for this study. The actual number of households finally interviewed, however, was considerably more than that since a certain proportion of accidents involved two or more persons residing at different addresses.

The Motor Vehicle Accident Report Form. After the sample was drawn, the North Carolina Department of Motor Vehicles was requested to provide the Division of Health Interview Statistics with a copy of the original accident report form for each accident falling in the sample. The report forms contained the name and address of the driver(s) involved regardless of whether he was injured and the name and address of all other persons in the accident who were either injured or killed. The record also contained a classification of the type of injury each injured person sustained and specific details of the accident. The three injury classifications were as follows:

1. A type A injury was described as a visible sign of injury, such as a bleeding wound or distorted limb, or where the person had to be carried from the accident scene;

2. A type B injury included other visible injuries or bruises, abrasions, swelling, limping and so forth;

3. A type C injury involved no visible sign of injury but the person experienced momentary unconsciousness or complained of pain.

The Field Operation. Once the motor vehicle accident report forms were received, the addresses of all persons residing in the three-county area were abstracted from the record. These sample addresses were then grouped into interview segments according to geographical proximity to one another. Prior to the actual interview, an advance letter was sent to each sample address informing the residents that they would be contacted by an interviewer from the U. S. Public Health Service who would ask them questions about the health of their family and other health-related items.

Interviewing was carried out by nine interviewers from the Health Interview Survey field staff over an eleven-week period, from February 20 through May 5, 1967. The interviewers received their training from staff members of the Division of Health Interview Statistics. The questionnaire the interviewers used for this study was a substantially shortened version of the schedule implemented in the ongoing program of the Health Interview Survey in 1968. In addition to the basic questionnaire, a motor vehicle accident supplement was completed for each reported accident; the supplement contained the detailed questions about the types of injuries sustained and other particulars about the accident.

As can be seen from Table A, an attempt was made to interview 939 sample households; of this number, 809 households were finally interviewed. Since the sample for this study was selected from records representing accidents occurring as much as a year preceding the interview, it was expected that some proportion of sample persons would be lost because they had moved from the original address. To minimize this kind of loss, however, a follow-up procedure was initiated:

> At the close of each interview, several questions were asked to obtain the name, and, if possible, the present address of any person who had resided in that household at any time during the past 12 months but who was not living there at the time of the interview. These questions had to be asked in each household since the interviewers, in most instances, were not given the name of the sample person, and, consequently, did not know when they were interviewing a sample family. An attempt

was eventually made to interview those sample persons reported to have moved from the original address provided: (1) they still lived within the three-county area, and (2) the new address obtained at the original household contained sufficient information to locate the person.

Table A. Number and percent distribution of completed interviews of sample and non-sample households by place of interview and number and percent distribution of non-interviews.

Interview Status	Number	Percent Distributions		
Total households	939	100.0		
Completed interviews	809	86.2	100.0	
Households yielding sample person:	640	68.2	79.1	100.0
at original address	625	66.6	77.3	97.7
at follow-up address	15	1.6	1.8	2.3
Households not yielding sample person:	169	18.0	20.9	100.0
Interview completed at original address	165	17.6	20.4	97.6
Interview completed at follow-up address	4	.4	.5	2.4
Non-interviews	130	13.8		

Of the 640 completed interviews (table A) resulting in a sample person, 15, or 2.3 percent, were the result of interviews conducted at a follow-up address. This percentage seems small when compared with the 625, or 97.7 percent, of the households yielding a sample person at the original address. However, the additional effort that went into locating these few sample persons seems worthwhile when considering that of the 19 interviews conducted at a follow-up address, slightly over three-fourths of them yielded a sample person.

About 14 percent of the households were never interviewed. This compares favorably with the percentage of type A, B, and C non-interviews in the ongoing Health Interview Survey. Table B shows a breakdown of these non-interviewed households by reason for non-interview.

The largest single factor contributing to the over-all non-interview rate was that in 23 households it was learned prior to the interview that the sample person no longer lived at the sample address. Another 17 households were never located by the interviewers. The problem of locating households occurred because, in some instances, the accident record contained an inaccurate or incomplete address. In this study, interviewers had the most difficulty finding households with rural addresses. The local Post Offices, however, provided some help in locating these households, minimizing the number of addresses that could not be located.

For the most part, the two types of non-interviews described above do not occur with any great frequency in the Health Interview Survey, and may explain why the non-interview rate in this study was higher than that found in the ongoing Survey. The percentage distributions for the other types of non-interviews, such as refusals, temporary absences, and demolished residences, were quite similar in the special study and the Survey.

Table B. Number and percent distribution of completed interviews and non-interviews by reason for non-interview.

Interview Status	Number	Percent Distribution	
Total households	939	100.0	
Completed interviews	809	86.2	
Non-interviews	130	13.8	100.0
Type A	51	5.4	39.2
Refusal	18	1.9	13.8
Not at home	22	2.3	16.9
Other	11	1.2	8.5
Type B	27	2.9	20.8
Vacant	18	1.9	13.8
Other	9	1.0	6.9
Type C	50	5.3	38.5
Not sample household	23	2.4	17.7
Could not locate house	17	1.8	13.1
Other	10	1.1	7.7
Non-interview status unknown	2	.2	1.5

III. Analysis of Data

The analysis of the record case study is based on the data in tables 1-8, and is discussed below. The solution to the problem of determining the optimum recall period will be treated in a later section of this report.

Table 1 shows that a total of 590 sample persons were interviewed. A sample person is defined as any person listed on the motor vehicle record who resided within the three-county area at the time of the accident. This includes all drivers, whether injured or not as indicated on the record, and all injured passengers. Also, any person reporting an injury, regardless of his injury status on the accident report form, is defined as a sample person.

Other facts of interest which are apparent from the data in table 1 are:

1. Eighty-two sample persons, or 13.9 percent of the total sample persons interviewed, did not report the accident.

2. The non-reporting of accidents increases as the time between the date of accident and interview increases. The non-reporting ranges from 3.4 percent for less than three months to a maximum of 27.3 percent for the interval of nine-twelve months. The obvious reason for this trend results from an increased inability to recall the occurrence of a motor vehicle accident as the time between the date of accident and the date of interview increases.

Of the 590 sample persons interviewed, the motor vehicle record indicates that 377 persons, or 63.9 percent, were injured (table 2). There are several points of interest evident from this table:

1. For the recall period of less than three months, 87.3 percent of the 377 injured persons interviewed reported the injury sustained in the accident. This compares with 78.8 percent for a recall period of less than six months, and 75.1 percent for less than twelve months.

2. Fifty-one sample persons, or 13.5 percent of the sample persons reported as injured on the motor vehicle record, reported the occurrence of the accident but did not report the injury. It is possible that the injury classification on the record may not be absolutely correct, and that not all of the 51 sample persons who reported the accident but failed to report the injury actually sustained an injury in the accident. However, since the record is being used as a criterion to estimate the ability of a respondent to report motor vehicle injuries, the bias introduced by inaccuracies in the record must be accepted. If the assumption is not made that the record is correct, no valid foundation exists on which the determination of the optimum recall period can be made.

3. Forty-three persons, or 11.4 percent of the sample persons classified on the record as being injured, did not report the sample accident. This percentage increased as the recall period became longer.

It can be seen from table 3 that, of the 590 sample persons who were involved in accidents and were interviewed, 213, or 36.1 percent, were reported as not injured on the motor vehicle record:

1. Thirty-four sample persons, or 16.0 percent of the 213 persons in this group, reported an injury when, in fact, the record indicated no injury. Most of these injuries were reported within a six-month recall period. This reporting trend may indicate that these injuries were minor, and were less likely to be reported as the recall period was extended beyond six months.

2. Thirty-nine persons, or 18.8 percent of the noninjured sample persons, did not report the accident. When this percentage is compared with the 11.4 percent of the injured persons who did not report the accident (shown in table 2), it seems that a respondent is more likely to report an accident if he received an injury in the accident.

The reporting of the accident and injury in the interview, by type of injury received (as indicated on the motor vehicle record), is shown in table 4. The following statistics are of importance:

1. For the recall period of less than twelve months, 85.5 percent of the type A injuries were reported, as compared to 67.8 percent type B, and 67.2 percent type C injuries (for definitions of type A, B, and C injuries, see page 2). This difference is significant, and can be interpreted as a result of the degree of severity of injury which is inherent in the definition of type A, B, and C injuries. This trend is also apparent for recall periods of three months and six months.

2. The reporting in the interview of type B and C injuries appears equally good. This similarity was unexpected, since type B injuries are, by definition, more severe than type C injuries.

3. For the recall period of less than twelve months, 10.7 percent of the sample persons who incurred type A injuries did not report the accident, as compared to 13.3 percent for type B, and 10.9 percent for type C injuries. These percentages indicate that the reporting of the accident is independent of the type of injury received. However, as indicated in point one above, the reporting of the injury itself is dependent on the type of injury received.

In this study, sample persons were classified into two response groups according to the following criteria. A sample person was classified as a self respondent if he or some other person(s) involved in the accident participated in the interview. If this condition is not met, the sample person is considered as having a proxy respondent.

The reporting and non-reporting of the accident and injury are shown by respondent status in table 5. Of particular interest in table 5 is the 3.9 percent of all self respondents, compared to 11.7 percent of the proxy respondents, who did not report the accident when the sample person was injured. Also, since self respondents are generally able to report most events more accurately than proxy respondents, it was surprising that 6.9 percent of self respondents, compared to only 4.3 percent proxy respondents, reported an injury for the sample person when the accident record indicated none. This reporting pattern probably occurred because a few sample persons who were not classified on the accident record as being injured actually received minor injuries, rather than because proxy respondents reported this item more accurately than did self respondents. Reporting differences for self and proxy respondents for the other categories in table 5 are small.

Inaccuracies in reporting the date of accident (by time interval) among sample persons reporting the accident are shown in table 6. The following points are of interest:

1. From table 1 it can be seen that, according to the record, 119 sample persons had an accident which occurred within a three-month period prior to interview. Of this number, 115 persons, or 96.6 percent, reported the accident. Of these 115 sample persons, 6 persons, or 5.2 percent, reported the accident as occurring in the interval three-six months prior to interview. This underreporting for the interval less than three months is counterbalanced by the reporting of 16 sample persons, or 8.5 percent of the sample persons, who had an accident in the interval three-six months prior to interview, but who reported in the interview that the accident occurred less than three months prior to the time of interview.

2. For the recall period less than six months, 5 persons, or 1.7 percent of the sample persons reporting the accident, reported it as occurring in the interval six-nine months prior to interview. This compares with 18 sample persons, or 8.7 percent, of the sample persons who reported the accident as occurring in the interval less than six months, when, according to the record, the accident occurred six-twelve months prior to interview.

3. Due to delays in interviewing, 59 sample persons were interviewed more than twelve months after the date of the accident. Of this number, 5 sample persons, or 8.5 percent of the 59 sample persons, reported the accident as occurring within the past twelve months.

4. The over-all pattern indicates that a certain proportion of the people who reported the sample accident, reported it as occurring earlier than the actual date of the accident. This phenomenon occurs at a slightly higher rate than the proportion of people who reported the occurrence of the accident on a date later than when it actually occurred. The net difference appears insignificant when examined for the three recall periods of less than three months, less than six months, and less than twelve months. For this reason, analysis of the optimum recall period will not be based on the bias in reporting of the date of the accident as shown in this table.

Information on the completeness of reporting of the accidents in all sample households is shown in tables 7 and 8, by interval since the occurrence of the accident and by respondent status:

1. Table 7 shows that 532 sample households resulted in an interview which yielded a sample person. The 532 households yielded 590 sample persons (table 1).

2. In 79 sample households, or 14.8 percent of the 532 households, the accident was not reported. This proportion compares with 13.9 percent of all sample persons who did not report the accident (table 3). Nonreporting of the accident increases as the recall period increases. The percentage ranges from 2.8 percent for recall of less than three months to 30.7 percent for the interval of 9-12 months.

3. Of the 532 sampled households, 287, or 53.9 percent, were self-responding households (table 8). A household is defined as self-responding if at least one person who responded in the interview was also in the accident, whether or not he was injured. If a household could not be classified in the self-responding category, then the household was defined in the proxy category. Of the 287 self-responding households, 29, or 10.1 percent, did not report the accident. This compares with 20.4 percent in the proxy-responding households. This difference is significant, and indicates the magnitude of bias which might result when information is obtained from a proxy respondent versus a self respondent.

IV. Determining the Optimum Recall Period

The National Center for Health Statistics has collected motor vehicle injury data in its National Health Interview Survey for the year 1968. The question asked of each respondent was: During the past twelve months, have you been in a motor vehicle accident, either as a driver, passenger, or a pedestrian? The data from this question have been processed and are available for analysis. National estimates of persons injured in moving motor vehicle accidents, as well as information about factors relating to the accident, are to be published. The purpose of the record case study is to assist in the determination of the recall period to be used in the interview survey that will give the most reliable estimate of P_{12}, the true proportion of motor vehicle injuries which have occurred in the United States during 1968. The concept and definition of a recall period have been discussed in the Analysis of Data section. The procedure for estimating the proportion of motor vehicle injuries which have occurred in the United States during 1968 is directly related to the recall period selected.

An example will best illustrate this relationship. If a less than three-month recall period is selected for estimating the total number of motor vehicle injuries occurring within the year, the procedure would be to estimate the total number of injuries occurring within a three-month interval and inflate this estimate to represent the total number of motor vehicle injuries occurring within the year. A reported injury in the Health Interview Survey is within a three-month interval if the respondent reported the injury as occurring within the three months prior to the date of interview. If the respondent reported the injury as occurring more than three months prior to date of interview, this injury would not be inflated in the estimation of the total number of motor vehicle injuries. A similar definition would hold for any other recall period period. The recall periods which will be considered in this analysis are: less than three months, less than six months, and less than twelve months.

For any recall period, there are two components of precision which must be carefully examined. The first of these is the variance of the estimator $\hat{P}_{12}$, and the second is the biasness of $\hat{P}_{12}$. $\hat{P}_{12}$ is the estimated proportion, or rate, of motor vehicle injuries occurring in the United States during the year 1968. There are two properties of variance and bias which are of importance when considering the three recall periods:

1. The variance of $\hat{P}_{12}$ decreases when the longer recall period is used. That is, a 12-month recall period results in a smaller variance for $\hat{P}_{12}$ when compared with the variance which results from using a six-month, or a three-month recall period.

2. The bias of $\hat{P}_{12}$ increases when the longer recall period is used. Bias is measured by the proportion of people who fail to report a motor vehicle injury. Bias increases since the ability of a respondent to recall a motor vehicle injury decreases as the recall period is lengthened.

The technique for determining the optimum recall period consists of the selection of the recall period which minimizes the sum of the variance component of $\hat{P}_{12}$ and the square of the bias of $\hat{P}_{12}$. In statistical terms, this method of optimization is referred to as the minimum mean-squared error, MSE.

Estimates of Variance, Mean-squared Error and Relative Root Mean-square Error. The sample size, estimated probability of injury, variance based on assumption of independence and non-independence, and the estimated bias of $\hat{P}_{12}$ squared are shown in table 9. The subscripts 3, 6, and 12 refer to three-month, six-month, and twelve-month recall periods, respectively. The following comments and explanations are needed:

1. The sample size N = 134,000. The sample size represents the estimated total number of people interviewed in the National Health Interview Survey in 1968, and remains constant for each of the three recall periods.

2. The true probability of a person receiving a motor vehicle injury for the entire year 1968 is denoted by P_{12}. It is assumed that this probability is uniform over the twelve-month period, and hence: $P_3 = \frac{1}{4}P_{12}$, and $P_6 = \frac{1}{2}P_{12}$, where P_3 and P_6 denote the probability of a person receiving a motor vehicle injury in a three-month and six-month time interval, respectively.

3. The variance of the estimated probability of injury in the past twelve months is shown for each of the recall periods. The variance which will be used in the analysis is the variance due to lack of independence. Independence is not satisfied in the National Health Survey, since the basic sampling unit is a household. That is, all of the respondents in a household would tend either to report or not report the accident and injuries. The variance of $\hat{P}_{12}$ due to lack of independence is expected to be at least twice as large as the variance of $\hat{P}_{12}$ if independence could be assumed.

4. The estimated bias of $\hat{P}_{12}$ squared is shown for each recall period. The factors K_3, K_6, and K_{12} are estimates of the bias in reporting motor vehicle injuries for three-month, six-month, and twelve-month recall periods, respectively. Estimates of the bias components from table 2 are: $K_3 = .127$, $K_6 = .212$, and $K_{12} = .249$. These estimates are based on the proportion of people who were reported as being injured on the motor vehicle record but failed to report the injury when interviewed.

5. The MSE of $\hat{P}_{12}$ by definition is equal to the variance of $\hat{P}_{12}$ plus the square of the bias of $\hat{P}_{12}$.[2] From table 9, the MSE of $\hat{P}_{12}$ can be determined for each recall period. Variance due to lack of independence is used in the MSE formula.

6. The RRMSE of $\hat{P}_{12}$ can be determined for each recall period from the formula:

$$\text{RRMSE}(\hat{P}_{12}) = \frac{\sqrt{\text{MSE}(\hat{P}_{12})}}{P_{12}} \times 100\%$$

It should be realized that the recall period which results in the minimum MSE of $\hat{P}_{12}$ will also result in the minimum RRMSE of $\hat{P}_{12}$. This can be seen be examining the RRMSE formula. The RRMSE shows the error of the estimate $\hat{P}_{12}$ as a percentage of the true proportion, P_{12}. In addition to selecting the recall period which gives the minimum RRMSE of $\hat{P}_{12}$, a further requirement is that the RRMSE of $\hat{P}_{12}$ for this recall period shall not exceed 25 percent. A RRMSE of 25 percent or less is considered an acceptable level for showing estimates of proportions or totals.

7. The variance component, X_1, of the RRMSE of $\hat{P}_{12}$ is the relative standard error of $\hat{P}_{12}$, or

$$X_1 = \frac{\sqrt{Var(\hat{P}_{12})}}{P_{12}} \times 100\%$$

It should be realized that the variance component is identical to the RRMSE of $\hat{P}_{12}$ if $\hat{P}_{12}$ is an unbiased estimator of P_{12}.

8. The bias component, X_2, of the RRMSE of $\hat{P}_{12}$ is the difference between the RRMSE of $\hat{P}_{12}$ and the variance component, or

$$X_2 = RRMSE(\hat{P}_{12}) - \frac{\sqrt{Var(\hat{P}_{12})}}{P_{12}} \times 100\%$$

In order to determine which recall period results in a minimum RRMSE from the RRMSE equation, it can be seen that we must assume a value for P_{12}. Data collected by the National Center for Health Statistics for the period July 1966-June 1967 show that an estimated 3.5 million persons, or a rate of 1.8 persons per 100, were injured in moving motor vehicle accidents. An estimate of P_{12} based on these data is 1.8×10^{-2}. Since it is desirable to show not only the estimated proportion of people injured, but also a categorization of this proportion by such characteristics as age, sex, driver status, residence, region, severity of accident, and possibly other variables, it is necessary to take these into consideration in the methodology, because an estimate of P_{12} based on these characteristics would be much smaller than 1.8×10^{-2}. For this reason, the optimum recall period is shown as a function of P_{12}. The following inequalities are solved for P_{12}.

1.1 $\quad RRMSE(4\hat{P}_3) \le RRMSE(2\hat{P}_6)$

1.2 $\quad RRMSE(4\hat{P}_3) \le RRMSE(2\hat{P}_{12})$

1.3 $\quad RRMSE(2\hat{P}_6) \le RRMSE(\hat{P}_{12})$

Let the solution of equation 1.1 for P_{12} be P'_{12}. This implies that a three-month recall period results in a smaller RRMSE, when compared to a six-month recall for all values of P_{12} less than or equal to P'_{12}. A similar interpretation holds in equation 1.2, comparison of three-month recall to a twelve-month recall; and equation 1.3, which compares a six-month recall to a twelve-month recall. By simultaneously considering equations 1.1, 1.2, and 1.3, and their solutions, a graph can be constructed showing the values of P_{12} which result in a minimum RRMSE for each of the recall periods. However, solutions in terms of P_{12} have very little intuitive meaning. For this reason solutions are shown in terms of T_{12}, where T_{12} is the population size of injured persons which results when P_{12} is inflated to represent the total United States population, that is, $T_{12} = 200 \times 10^6 \times P_{12}$.

Graph 1. The recall period resulting in the minimum RRMSE for specific population sizes of injured persons.

Twelve-month Recall Period	Six-month Recall Period	Three-month Recall Period	
174,000	195,000	207,000	T_{12}

Graph 1 above shows the population sizes of injured persons which result in a minimum RRMSE for each of the recall periods. The following statistics are of interest:

1. Graph 1 shows that a twelve-month recall period results in a minimum RRMSE for estimates on specific injured populations of size less than 174,000. For estimates ranging from size 174,000 to 207,000, a six-month recall period results in a smaller RRMSE when compared to a twelve-month recall period of injured populations of size greater than 195,000. For estimates larger than 207,000, a three-month recall period yields the minimum RRMSE over both the six-month and twelve-month recall periods.

2. A twelve-month recall period yields the minimum RRMSE for estimates on small population sizes. As the population size increases, the six-month recall period becomes optimum over the twelve-month recall period. This occurs at a population of size 154,000. Eventually, the population size increases to a point (207,000) where the three-month recall yields the minimum RRMSE.

However, graph 1, above, does not show the actual value of the RRMSE, but only the population sizes for which each recall period yields the minimum RRMSE. Table 10 shows the value of the RRMSE, the variance component, and the bias component for each recall period as the population varies in size from 25 thousand to 5 million. Based on data from this table and graph 1, the optimum recall period for estimating the total number of persons injured in motor vehicle accidents is the recall period of less than three months. The following reasons support this decision:

1. The RRMSE of estimates larger than 207,000 is a minimum for the less than three-month recall period. As the size of the estimates increases, table 10 shows that the RRMSE based on a recall period of 12 months decreases slightly from 27.7 percent to 25.0 percent. For a less than six-month recall period, this decrease is from 27.3 percent to 21.5 percent. The largest decrease occurs in the three-month recall period, where the RRMSE declines from a level of 27.5 percent to 13.6 percent.

2. Estimates of greatest interest are for populations of size greater than 207,000. Indeed, the single most important estimate is the

total number of moving motor vehicle injuries, which is estimated to be nearly 4 million. The RRMSE for an estimate of 4 million is 25.0, 21.5, and 13.8 percent for a twelve-month, six-month, and three-month recall period, respectively. The contrast in these three percentages led to the selection of a recall period covering the three-month interval preceding the week of interview.

3. Data collected in the Health Interview Survey for the period July to December 1967 have been processed and estimates of the total number of persons injured in motor vehicle accidents within the year have been made using each of the recall periods. The estimated total number of persons injured is 3.2, 2.7, and 2.4 million based on three-month, six-month, and twelve-month recall periods, respectively. From a comparison of these estimates, it seems that the bias component of the RRMSE, which is a function of the ability of a respondent to recall a motor vehicle injury, increases over time at a rate greater than estimated from this methodology study. Hence, it appears that the results of this study, which led to the selection of a three-month recall period, are conservative.

4. The variance component and bias component have certain effects on the value of the RRMSE. As the estimated number of persons injured increases, the variance component of the RRMSE decreases, the bias component increases, and the RRMSE decreases (figure 2).

REFERENCES

[1] National Center for Health Statistics, Current Estimates from the Health Interview Survey, PHS Publication No. 1000, Series 10, No. 43, p. 3, Washington, D.C., January 1968.

[2] Graybill, F. A., and Mood, A. M. (1963), Introduction to the Theory of Statistics, New York, McGraw-Hill.

Table 1. Reporting and non-reporting of sample accident for all sample persons interviewed.

Time Lapse Between Date of Accident and Date of Interview	All Sample Persons: Total		Reported Sample Accident		Did not report Sample Accident	
	Frequency	Percent Dist.	Frequency	Ratio Dist.	Frequency	Ratio Dist.
Total	590	100.0	508	86.1	82	13.9
Less than 3 months	119	20.2	115	96.6	4	3.4
3 - 6 months	209	35.4	187	89.5	22	10.5
6 - 9 months	119	20.2	102	85.7	17	14.3
9 - 12 months	143	24.2	104	72.7	39	27.3
Less than 6 months	328	55.6	302	92.1	26	7.9
6 - 12 months	262	44.4	206	78.6	56	21.4
Less than 12 months	590	100.0	508	86.1	82	13.9

Table 2. Reporting and non-reporting of sample accident and injury for all sample persons who were injured on Motor Vehicle Record.

Time Lapse Between Date of Accident and Date of Interview	Sample Person Injured on Motor Vehicle Record									
	Total		Sample Accident Reported						Sample Accident Not Reported	
					Reported Injury		Did not Report Injury			
	Frequency	Percent Dist.	Frequency	Ratio Dist.	Frequency	Ratio Dist.	Frequency	Ratio Dist.	Frequency	Ratio Dist.
Total	377	100.0	334	88.6	283	75.1	51	13.5	43	11.4
Less than 3 months	71	18.8	70	98.6	62	87.3	8	11.3	1	1.4
3-6 months	141	37.4	127	90.1	105	74.5	22	15.6	14	9.9
6-9 months	71	18.8	64	90.1	57	80.3	7	9.9	7	9.9
9-12 months	94	24.9	73	77.7	59	62.8	14	14.9	21	22.3
Less than 6 months	212	56.2	197	92.9	167	78.8	30	14.2	15	7.1
6-12 months	165	43.8	137	83.0	116	70.3	21	12.7	28	17.0
Less than 12 months	377	100.0	334	88.6	283	75.1	51	13.5	43	11.4

Table 3. Reporting and non-reporting of sample accident and injury for all sample persons not injured on Motor Vehicle Record.

Time Lapse Between Date of Accident and Date of Interview	Sample Person Not Injured on Motor Vehicle Record									
	Total		Sample Accident Reported						Sample Accident Not Reported	
					Did Not Report Injury		Did Report Injury			
	Frequency	Percent Dist.	Frequency	Ratio Dist.	Frequency	Ratio Dist.	Frequency	Ratio Dist.	Frequency	Ratio Dist
Total	213	100.0	174	81.2	140	65.7	34	16.0	39	18.8
Less than 3 months	48	22.5	45	93.8	30	62.5	15	31.3	3	6.2
3-6 months	68	31.9	60	88.2	50	73.5	10	14.7	8	11.8
6-9 months	48	22.5	38	79.2	34	70.8	4	8.3	10	21.8
9-12 months	49	23.0	31	63.3	26	53.1	5	10.2	18	36.7
Less than 6 months	116	54.5	105	90.5	80	69.0	25	21.6	11	9.5
6-12 months	97	45.5	69	71.1	60	61.9	9	9.3	28	28.9
Less than 12 months	213	100.0	174	81.2	140	65.7	34	16.0	39	18.8

Table 4. Reporting of accident and injury by type of injury on Motor Vehicle Record.

Time Lapse Between Date of Accident and Date of Interview	Injury Classification on Motor Vehicle Record: Type A Total	Type A Reported Injury	Type A Reported Accident Only	Type A Did Not Report Accident	Type B Total	Type B Reported Injury	Type B Reported Accident Only	Type B Did Not Report Accident	Type C Total	Type C Reported Injury	Type C Reported Accident Only	Type C Did Not Report Accident
Less than 3 months												
Number	29	27	2	-	13	11	2	-	29	24	4	1
Percent Dist.	100.0	93.1	6.9	-	100.0	84.6	15.4	-	100.0	82.6	13.7	3.7
3-6 months												
Number	59	53	1	5	34	23	7	4	48	29	14	5
Percent Dist.	100.0	89.8	1.7	8.5	100.0	67.6	20.3	11.8	100.0	60.4	29.2	10.4
6-9 months												
Number	27	24	1	2	15	13	1	2	28	20	5	3
Percent Dist.	100.0	88.9	3.7	8.4	100.0	81.3	6.2	12.5	100.0	71.4	17.9	10.7
9-12 months												
Number	44	32	2	10	27	14	7	6	23	13	5	5
Percent Dist.	100.0	72.7	4.5	22.8	100.0	51.8	26.0	22.2	100.0	56.6	21.7	21.7
Less than 6 months												
Number	88	80	3	5	47	34	9	4	77	53	18	6
Percent Dist.	100.0	90.9	3.4	5.7	100.0	72.3	19.1	8.5	100.0	68.8	23.8	7.4
6-12 months												
Number	71	56	3	12	43	27	8	8	51	33	10	8
Percent Dist.	100.0	78.9	4.2	16.9	100.0	62.8	18.6	18.6	100.0	64.7	19.6	15.7
Less than 12 months												
Number	159	136	6	17	90	61	17	12	128	86	28	14
Percent Dist.	100.0	85.5	3.8	10.7	100.0	67.8	18.9	13.3	100.0	67.2	21.9	10.9

Table 5. Number and percent distribution of reporting and non-reporting of accident and injury by respondent status.

Respondent Status	Total	Sample Person Did Not Report Accident		Sample Person Reported Accident Only		Sample Person Reported Injury	
		Injured on Record	Not Injured on Record	Injured on Record	Not Injured on Record	Injured on Record	Not Injured on Record
Total Persons	590	43	39	51	140	283	34
Self Respondents							
Number	333	13	18	29	80	170	23
Percent Dist.	100.0	3.9	5.4	8.7	24.0	51.1	6.9
Proxy Respondents							
Number	257	30	21	22	60	113	11
Percent Dist.	100.0	11.7	8.2	8.5	23.3	44.0	4.3

Table 6. Error in reporting date of accident by time interval for all sample persons reporting sample accident.

Time Lapse Between Date of Accident and Date of Interview	Total	Sample Accident Reported as Occurring in the Following Time Interval from Date of Interview				
		Less than 3 months	3-6 months	6-9 months	9-12 months	DK Date Given
Total*	508	125	189	100	90	4
Less than 3 months	115	109	6	-	-	-
Percent distribution	100.0	94.8	5.2	-	-	-
3 - 6 months	187	16	165	5	-	1
Percent distribution	100.0	8.5	88.2	2.7	-	.6
6 - 9 months	102	-	14	83	3	2
Percent distribution	100.0	-	13.7	81.4	2.9	2.0
9 - 12 months	104	-	4	12	87	1
Percent distribution	100.0	-	3.8	11.5	83.7	1.0
Less than 6 months	302	125	171	5	-	1
Percent distribution	100.0	41.4	56.6	1.7	-	.3
6 - 12 months	206	-	18	95	90	3
Percent distribution	100.0	-	8.7	46.1	43.7	1.5
12+ months	59	-	-	-	5	
Percent distribution	100.0	-	-	-	8.5	

* Excludes 12+ months

Table 7. Comparison of the date of accident on record and questionnaire for all households interviewed containing 1+ sample persons.

Time Lapse Between Date of Accident and Date of Interview	All Sample Households					
	Total		Reported Sample Accident		Did not report Sample Accident	
	Frequency	Percent Dist.	Frequency	Ratio Dist.	Frequency	Ratio Dist.
Total	532	100.0	453	85.2	79	14.8
Less than 3 months	106	19.9	103	97.2	3	2.8
3 - 6 months	198	37.2	177	89.4	21	10.6
6 - 9 months	101	19.0	85	84.2	16	15.8
9 - 12 months	127	23.9	88	69.3	39	30.7
Less than 6 months	304	57.1	280	92.1	24	7.9
6 - 12 months	228	42.9	173	75.9	55	24.1
Less than 12 months	532	100.0	508	85.2	79	14.8

Table 8. Comparison of the date of accident on record and questionnaire for all households interviewed by respondent status.

Time Lapse Between Date of Accident and Date of Interview	All Sample Self-responding Households						All Sample Proxy-responding Households					
	Total		Reported Sample Accident		Did not report Sample Accident		Total		Reported Sample Accident		Did not report Sample Accident	
	Frequency	Percent Dist.	Frequency	Ratio Dist.	Frequency	Ratio Dist.	Frequency	Percent Dist.	Frequency	Ratio Dist.	Frequency	Ratio Dist.
Total	287	100.0	258	89.9	29	10.1	245	100.0	195	79.6	50	20.4
Less than 3 months	54	18.8	51	94.4	3	5.6	52	21.2	52	100.0	0	0.0
3 - 6 months	107	37.3	102	95.3	5	4.7	91	37.1	75	82.4	16	17.6
6 - 9 months	55	19.2	51	92.7	4	7.3	46	18.8	34	74.0	12	26.0
9 - 12 months	71	24.7	54	76.1	17	23.9	56	22.9	34	60.7	22	39.3
Less than 6 months	161	56.1	153	95.0	8	5.0	143	58.4	127	88.8	16	11.2
6 - 12 months	126	43.9	105	83.3	21	16.7	102	41.6	68	66.6	34	33.3
Less than 12 months	287	100.0	258	89.9	29	10.1	245	100.0	195	79.6	50	20.4

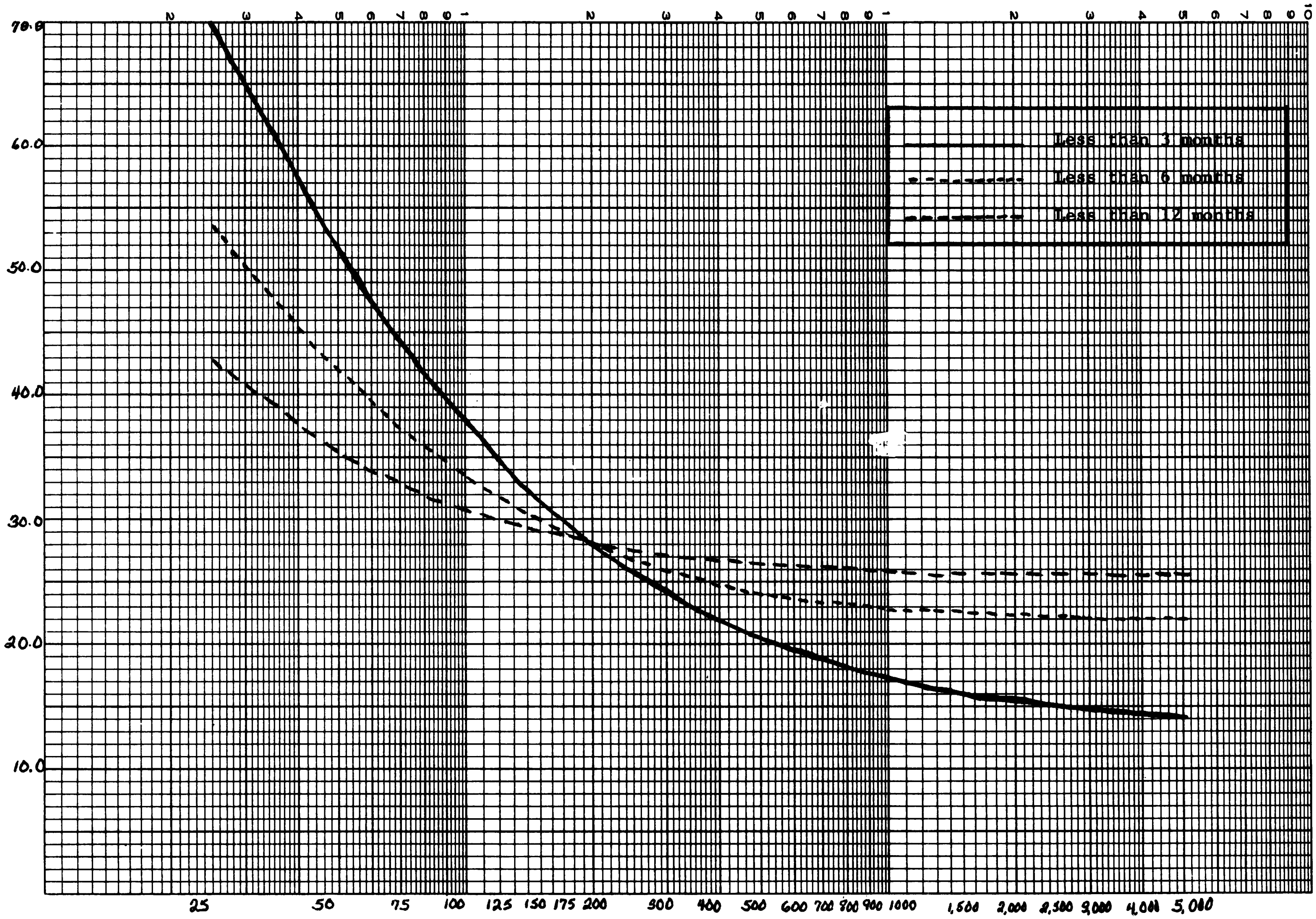
Table 9. Variance and bias of the estimated probability of injury by recall period.

	RECALL PERIOD		
	Less than 3 months	Less than 6 months	Less than 12 months
Sample size	$n = 134 \times 10^3$	$n = 134 \times 10^3$	$n = 134 \times 10^3$
Probability of person being injured	$P_3 = \frac{1}{4}P_{12}$	$P_6 = \frac{1}{2}P_{12}$	P_{12}
Estimated probability of person being injured in past 12 months	$4\hat{P}_3$	$2\hat{P}_6$	$\hat{P}_{12}$
Variance of estimated probability of injury in past 12 months, based on assumption of independence	$VAR\ (4\hat{P}_3) = 16P_3 \cdot Q_3/n = \frac{4P_{12}(1-\frac{1}{4}P_{12})}{n}$	$VAR\ (2\hat{P}_6) = 4P_6 \cdot Q_6/n = \frac{2P_{12}(1-\frac{1}{2}P_{12})}{n}$	$VAR\ (\hat{P}_{12}) = P_{12} \cdot Q_{12}/n = \frac{P_{12}(1-P_{12})}{n}$
Variance due to lack of independence	$K \cdot VAR\ (4\hat{P}_3) \quad K \geq 2$	$K \cdot VAR\ (2\hat{P}_6) \quad K \geq 2$	$K \cdot VAR\ (\hat{P}_{12}) \quad K \geq 2$
Proportion of persons reported as injured on motor vehicle record who reported an injury when interviewed	$K_3 = .873$	$K_6 = .788$	$K_{12} = .751$
Estimated bias of $\hat{P}_{12}$ squared = $(\hat{P}_{12} - P_{12})^2$	$(4P_3-4K_3P_3)^2 = P_{12}^2(1-K_3)^2$	$(2P_6-2K_6P_6)^2 = P_{12}^2(1-K_6)^2$	$(P_{12}-K_{12}P_{12})^2 = P_{12}^2(1-K_{12})$

Table 10. Relative root mean square error, variance and bias components for selected population sizes of injured persons.

Total Persons (in thousands)	Less than 3 months			Less than 6 months			Less than 12 months		
	Relative Root Mean Square Error %	Variance Component %	Bias Component %	Relative Root Mean Square Error %	Variance Component %	Bias Component %	Relative Root Mean Square Error %	Variance Component %	Bias Component %
25 -----	70.3	69.1	1.2	53.2	48.8	4.4	42.6	34.6	8.0
50 -----	50.5	48.9	1.6	40.5	34.5	6.0	34.9	24.4	10.5
75 -----	41.9	39.9	2.0	35.3	28.2	7.1	31.9	19.9	12.0
100 -----	36.8	34.5	2.3	32.3	24.4	7.9	30.3	17.3	13.0
125 -----	33.4	30.9	2.5	30.4	21.8	8.6	29.3	15.4	13.9
150 -----	30.9	28.2	2.7	29.1	19.9	9.2	28.6	14.1	14.5
175 -----	29.0	26.1	2.9	28.1	18.5	9.6	28.1	13.1	15.0
200 -----	27.5	24.4	3.1	27.3	17.3	10.0	27.7	12.2	15.5
300 -----	23.6	19.9	3.7	25.5	14.1	11.4	26.8	10.0	16.8
400 -----	21.4	17.3	4.1	24.5	12.2	12.3	26.3	8.6	17.7
500 -----	20.0	15.5	4.5	23.8	10.9	12.9	26.1	7.7	18.4
600 -----	19.0	14.1	4.9	23.4	10.0	13.4	25.9	7.1	18.8
700 -----	18.2	13.0	5.2	23.1	9.2	13.9	25.7	6.5	19.2
800 -----	17.6	12.2	5.4	22.9	8.6	14.3	25.6	6.1	19.5
900 -----	17.1	11.5	5.6	22.7	8.1	14.6	25.6	5.8	19.8
1,000 -----	16.7	10.9	5.8	22.5	7.7	14.8	25.5	5.5	20.0
1,500 -----	15.5	8.9	6.6	22.1	6.3	15.8	25.3	4.5	20.8
2,000 -----	14.9	7.7	7.2	21.9	5.5	16.4	25.2	3.9	21.3
2,500 -----	14.5	6.9	7.6	21.8	4.9	16.9	25.1	3.4	21.7
3,000 -----	14.2	6.3	7.9	21.6	4.4	17.2	25.1	3.1	22.0
3,500 -----	14.0	5.8	8.2	21.6	4.1	17.5	25.0	2.9	22.1
4,000 -----	13.8	5.4	8.4	21.5	3.8	17.7	25.0	2.7	22.3
4,500 -----	13.7	5.1	8.6	21.5	3.6	17.9	25.0	2.6	22.4
5,000 -----	13.6	4.9	8.7	21.5	3.5	18.0	25.0	2.4	22.6

Figure 2. Relative root mean square for aggregates by recall period.



XV

CONTRIBUTED PAPERS V

EDUCATION AND OCCUPATION: MODELS, METHODS AND FINDINGS

Chairman, PATIENCE LAURIAT, Social Security Administration

A DEMOGRAPHIC MODEL OF STUDENT PROGRESSION

W. L. Li
Ohio State University*

During the past few years there has been considerable interest in constructing mathematical models for the study of the educational process. These models, according to their subjects, can be classified into three categories (See Wurtele, 1967). The first category may be called demographic models. They focus primarily on the educational system or some of its components, such as the flow of students, teacher-student ratio, etc. The second category are the econometric models which treat education as one of the several interrelated economic activities; educational institutions are viewed as producers of outputs that are employed by the different sectors of society. The third category of models deals with the learning outcomes of individual students, or group of students. The socio-psychological aspects of educational process are strongly emphasized.

This paper is concerned with the first category of educational models. It attempts to examine student progression in an educational system from the demographic point of view. So, the subject of educational process is treated in the aggregate, and the interdependencies of the educational system with other sectors of society is analytically disregarded.

A General Demographic Model

The subject of educational process has long been of great interest to demographers. And the demographic analysis of educational process has been a great contribution to the educational planners, who must continually estimate the size of future student enrollments at different levels of the educational structure. Examples are found in the work of the Census Bureau, which in the past years has provided a continuous projection of school enrollments (Census Bureau, 1963; Siegel, 1967). A systematic exposition of educational demography using the Census Bureau's statistics is shown in the publication of Folger and Nam (1967).

However, demographers are often blamed for their failure to make an accurate educational projection. It is sometimes complained that demographers have relied too much on the techniques of trend extrapolation. Besides, the reliability of school enrollment analysis seems to be dependent on the birth-death projection of the total population which itself may be inaccurate.

Let us begin with an examination of the general demographic methodology of projecting school population. It can be best summarized in the following equation (Stone, 1966):

$$s = n \cdot q \qquad (1)$$

where s is the total student enrollment; n is the population vector of each age group; q is also a vector which gives the age-specific enrollment rates.

One way to implement the population vector is to follow Leslie's matrix approach, as Stone suggests. The Census Bureau's projection technique, though slightly different, nevertheless, is more or less based on trend extrapolation. Similarly, the enrollment rates are generally computed as a linear projection of the trend in observed fall enrollment rates in the past years.

One of the main limitations of such model, as Correa points out (Correa, 1967:34), is that the projected educational enrollments closely reflect the differences in the population structure and they are inappropriate to be used for temporal or spatial comparisons. There is another important limitation of such a model. It is that the model fails to give sufficient attention to the underlying dynamics of the educational process. This weakness is similar to the econometricians' construct of labor force function, which yields very limited knowledge of how the size of the labor force is determined by population structure.

Structure of Educational System

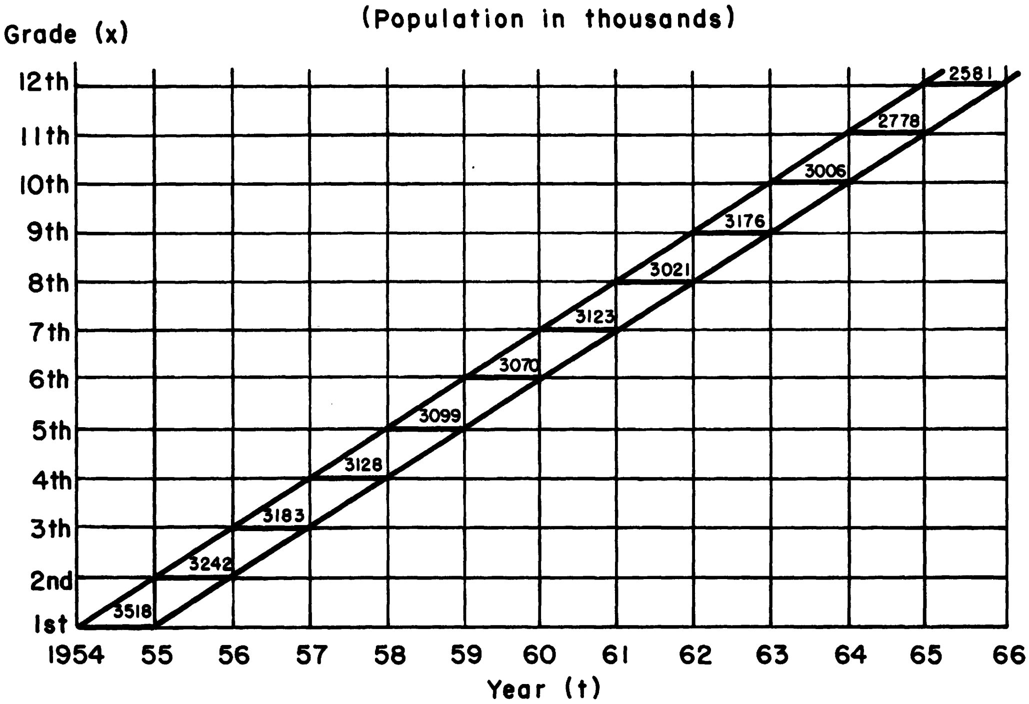
In order to reach a more realistic projection of student population, we propose to begin with an examination of the underlying mechanism of educational process. The major part of this paper attempts to assess the underlying dynamics of cohort student progression. Let us first consider the following Lexis diagram which represents the progression of a student cohort (Fig. 1).1/

The example shows that the student cohort first enters the educational system in 1954. With the increment of years, its size changes from the first grade to the last grade. Symbolically the size of the cohort may be denoted as $N(x,t)$, where $x = 1, 2, \ldots 12$ and $t = 1954, 55, \ldots$ It is obvious that the size tends to decrease over time in a population which is closed against immigration. If we take the initial size of the cohort as a basis, it can be shown (Fig. 2) that the decline of this cohort size is very much similar to a negative exponential distribution. It can be generalized as having the form of $N(x) = N(1) \cdot e^{-kt}$, where k is a constant term. For different cohorts, there will be different constant terms. It is possible to test empirically the variation of the terms for

* This research has been supported by the National Sciences Foundation under contract NSF-GS2630.

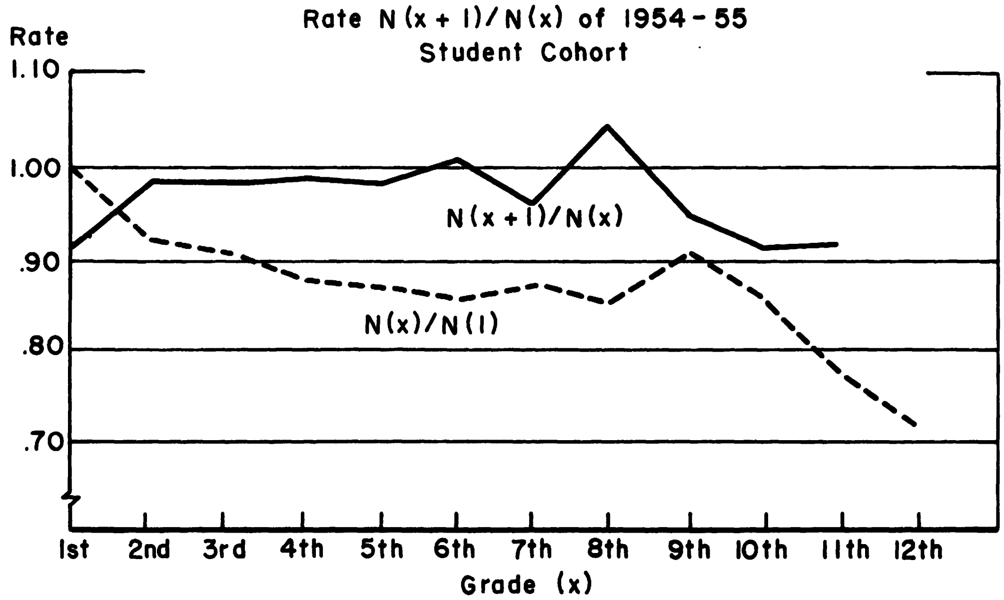
1/ An extensive application of Lexis diagram to demographic analysis can be found in Pressat (1961).

Figure 1
Progression of 1954-55 Student Cohort
(Population in thousands)



Source: Simon and Grant (1966: Table 28)

Figure 2
Survival Rate N(x)/N(1) and Retention
Rate N(x + 1)/N(x) of 1954-55
Student Cohort



cohorts at different points in time; so some interesting results may be shown.

Nevertheless, examining only the behavior of the constant terms can be misleading. Obviously the curve is not a smooth one. A more realistic approach is to analyze a student cohort's "retention rate" (Duncan, 1965:129).2/ Symbolically it is N(x+1, t+1)/N(x,t). Fig. 2 indicates that the rate does not show any tendency of monotonous decline. On the contrary, it increases from the first grade to the sixth grade. Only after the eighth grade does the rate begin to decline monotonously. This observation leads to the conclusion that the decline of a student cohort is not strictly comparable to the survival curve in the general human population.

This suggests that a student population of a given grade is not only affected by those separating forces such as mortality and dropout. Many other factors can also contribute to determine the cohort size. Probably the most important factor is the effect of failure (or repeaters) at each level of an educational system. The effect is analogous to that of a rolling snowball; the repeaters tend to increase the rate of retention as shown in Fig. 2. Some other less important factors can be immigration and the reenrollment of dropouts.

Normally these two forces work in the opposite direction: mortality and dropout on the one hand, failure and new entry on the other hand. If the separating force is predominant, the size of a student cohort tends to decrease, and, hence, the retention rate will be less than unity. However, the effects of failure and new entry can be so strong that the decreasing tendency is cancelled out, which is obviously the case in Fig. 2. The combined effect of mortality and dropout is called the "separation factor" (Stockwell and Nam, 1963).

An Educational Model from the Markov Process

Several simplifications must be made before we can construct a model which will take account of these underlying dynamics. We shall limit ourselves to the analysis of a closed population, an assumption which is not uncommon in demographic analysis. Furthermore, following a model of the Norwegian educational system (Thonstad, 1967), we assume that dropout is a one-time process; even though some dropouts re-enroll in schools in the later years, there are reasons to believe that the number may be too small to affect the total size of a school population. In other words, it is assumed that dropout is an absorbing state. Once an individual has left the educational system, he will not return.

2/ Here the concept of "retention rate" is used following the examples of Duncan (1965) and Stockwell and Nam (1963). However, the concept is used differently in the Office of Education's publications.

Based on these assumptions, it follows that the progression of a student cohort has four alternatives, namely, failure, progress, death and dropout. The flow distribution of the cohort is a vector, and so the total educational system can be contrived as an input-output matrix, shown in Table 1.3/ The figures in the matrix are purely illustrative. The first three rows represent the educational system up to the third grade; their row sums show the number of students in a given grade; their elements give the flow distribution of those students in that grade. The last three rows represent graduation (g), mortality (m) and dropout (d), respectively. Their row sums are zero, as they are assumed to be absorbing states.

TABLE 1

AN EXAMPLE OF STUDENT PROGRESSION

t1 \ t2	1st grade	2nd grade	3rd grade	Graduation	Mortality	Dropout	Sum
1st grade	75	415	0	0	5	5	500
2nd grade	0	45	391	0	5	9	450
3rd grade	0	0	20	360	4	16	400
Graduation	0	0	0	0	0	0	0
Mortality	0	0	0	0	0	0	0
Dropout	0	0	0	0	0	0	0
Sum	75	460	411	360	14	30	1350

It is obvious that such an input-output matrix can be readily formulated into a transition matrix, shown as follows:

	1	2	3	g	m	d
1	.15	.83	0	0	.01	.01
2	0	.10	.87	0	.01	.02
3	0	0	.05	.90	.01	.04
g	0	0	0	0	0	0
m	0	0	0	0	0	0
d	0	0	0	0	0	0

All elements in the matrix give the transitional probabilities from one state to another. Let us use A(i·j) to denote such a transitional matrix. If the matrix is treated as stationary, it is possible to obtain the grade distribution of a student population at the successive years through an application of the elementary Markov process principle. It involves the use of the following equation:

$$N^{t+1}_{(i)} = N^{t}_{(i)} \cdot A_{(i,j)} \qquad (2)$$

where the grade distribution of students at time t+1 is the result of multiplying the grade distribution of students at time t by the transitional matrix.

3/ Treating educational system as an input-output matrix has been shown in OECD (1967), Ch. 2.

This approach has been used extensively in many countries for the projection of future school enrollments. (See, for example, Zabrowski et al., 1967; Thonstad, 1967). It is not the purpose of our paper to repeat the same effort; however, we find that this approach can serve as a powerful analytical device. We are going to show that a model of student progression can be generated from this approach.

Let us take the example of projecting the 2nd grade school population at time t+1. According to the Markov process principle expressed in Equation (2), it shall be:

$$(3) \quad N^{t+1}_{(2)} = N^{t}_{(1)} \cdot A_{(1,2)} + N^{t}_{(2)} \cdot A_{(2,2)}$$

where $A_{(1,2)}$ is the progression rate from the 1st grade to the 2nd grade; $A_{(2,2)}$ is the failure rate for the 2nd grade students to remain in the same grade. What Equation (3) says is that the size of the student cohort at x+1th grade in the t+1th year is composed of those students of $\overline{N(x,t)}$ who can successfully progress to x+1th grade, and those students of N(x+1,t) who have to repeat in the same grade.

To extend this line of thinking, let us use some symbols to denote the following concepts:

N(x,t) = the student population at the xth grade in the tth year.

m(x) = the mortality rate of the xth grade.

d(x) = the dropout rate at the xth grade.

r(x) = the failure rate at the xth grade when students are supposed to progress to the x+1th grade.

p(x) = the progression rate for those who are in the xth grade and successfully progress to the x+1th grade.

Accordingly, a model can be generated from the aforementioned Markov-process consideration. The model, which is similar to that of equation (3), can be expressed as follows:

$$(4) \quad N(x+1, t+1) = N(x, t) \cdot p(x) + N(x+1, t) \cdot r(x+1)$$

Besides, it is seen from the transitional matrix that:

$$(5) \quad p(x) = 1 - (m(x) + d(x) + r(x))$$

So the model in Equation (4) becomes:

$$(6) \quad N(x+1,t+1) = N(x,t) \cdot (1 - (m(x) + d(x) + r(x))) + N(x+1,t) \cdot r(x+1)$$

A Conversion Technique

It appears that this model is not unique in the field of educational planning. As a matter of fact, most of the mathematical models of student progression are structurally more or less similar. However, they are different in the methods of implementing their models. The main effort of this paper, therefore, lies on the fact that we attempt to incorporate some of the existing demographic techniques into the implementation of our model.

Age is one of the basic demographic characteristics. It is "an invariant function of time and in this sense a fixed variable" (Duncan, 1965:123). So demographers are more inclined to measure social phenomenon by using age as an unit of measurement. This is very prevalent in the literature of educational demography. However, it is generally understood that the educational system is a structure of grade hierarchy. Consequently, the accomplishments of educational demographers are often ignored and their works are scarcely incorporated in the educational model-building. For instance, the work of Stockwell and Nam is cited by some educational planners (e.g., Correa, 1967; Dressel, 1967), and yet their ideas, particularly the measurement of dropout, have never been seriously employed. Here we attempt to show that some of the Stockwell and Nam's ideas can be readily applied to educational planning.

Stockwell and Nam's measurement of dropout is very similar to the measurement of retirement in labor force analysis (Wolfbein, 1949). It is a by-product of constructing the table of school life. First of all, there is a concept called "stationary school population," which is the result of multiplying age-specific enrollment rates by the stationary (life table) population.4/ Let us denote age as a, and the stationary school population at age a as N'(a). The ratio of the population at the successive ages, N'(a+1)/N'(a), shows the proportion of the life table population who remain in schools during a to a+1 interval. The complement of this proportion gives the propensity that the population at age a will not be enrolled in schools at age a+1. The causes of this "separation factor" appear to be mortality and dropout. So dropout rate can be obtained by operationally differentiating mortality rate from this separation factor.

The mortality rate in a life table is simply the ratio of deaths to the life table population at a given age, where deaths are those in a population who do not survive from age a to age a+1. We shall denote the mortality rate as m(a). Then the propensity of dropout at age a is measured by:5/

4/ For those who are not familiar with techniques of life table construction, a classical reference is: Dublin, L. I., A. J. Lotka, and M. Spiegelman, Length of Life, Ronald Press Co., N. Y., 1949.

5/ The formula is slightly different from what Stockwell and Nam present. They have adjusted the denominator (population) by a half of deaths. To the present writer, the adjustment seems to be unnecessary; however, it is not the major concern of this paper to discuss such a technical difference.

(7) $d(a) = 1 - N'(a+1)/N'(a) - m(a)$

Table 2 presents the rates of dropout and mortality by age for the United States school population in 1960. The mortality rate is constructed by assuming that the risk of death in the school population is the same as the general population. This is perhaps not too far from the truth. The dropout rate is the residual of substrating the mortality rate from the total separation factor, as Equation (7) indicates. It is shown that the dropout rate increases from 2 per thousand at ages 10-11 to 350 per thousand at ages 18-19, and then it decreases, with fluctuations, to 120 per thousand at ages 29-30+.

It appears that an empirical finding such as this is very consistent with our general knowledge of student's progression in an educational system. For instance, the peak of dropping out of school occurs around age 18, which is approximately the time when students have completed high school. This observation raises a further inquiry, namely, to what extent the dropout rates in successive ages are corresponding to the dropout rate in a successive grade?

The question is also of practical importance, as it is generally observed that in this country different statistical agencies tend to use different classifications in publishing their educational data: some are age-specific, whereas others are grade-specific. For this reason, it is analytically necessary to obtain a transformation matrix which can show a correspondence between age and sex characteristics, so that a manipulation of the matrix will be able to convert any demographic characteristic from an age vector into a grade vector.

Let us denote a school population at age a of grade x as $S(a,x)$. The total school population of all grades for age a will be $\sum_x S(a,x)$. It follows that there can be a matrix, T, with its elements as:

TABLE 2

MORTALITY AND DROPOUT AND SEPARATION RATES BY AGE
UNITED STATES, 1960

Age	Population		Enroll.	Separation Rate (per thousand)		
	Life Table	School	Ratio	Total	Death	Dropout
5	96,977	43,446	.448	0.7	0.7	-
6	96,913	80,825	.834	0.5	0.5	-
7	96,860	93,954	.970	0.5	0.5	-
8	96,815	94,685	.978	0.4	0.4	-
9	96,776	94,744	.980	0.4	0.4	-
10	96,741	94,709	.979	2.4	0.4	2.0
11	96,707	94,483	.977	2.4	0.4	2.0
12	96,671	94,254	.975	6.6	0.4	6.2
13	96,631	93,635	.969	17.0	0.5	16.5
14	96,583	92,044	.953	26.8	0.6	26.2
15	96,526	89,576	.928	69.6	0.7	68.9
16	96,457	83,339	.864	124.6	0.8	123.7
17	96,377	72,957	.757	333.5	0.9	332.6
18	96,287	48,625	.505	351.2	1.0	350.1
19	96,189	31,550	.328	281.3	1.1	280.2
20	96,085	22,676	.236	208.5	1.1	207.4
21	95,975	17,947	.187	237.7	1.2	236.5
22	95,859	11,887	.124	210.7	1.3	209.4
23	95,739	9,382	.098	154.1	1.3	152.9
24	95,618	7,936	.083	85.5	1.3	84.2
25	95,496	7,258	.076	93.3	1.3	92.0
26	95,374	6,581	.069	102.6	1.3	101.3
27	95,252	5,906	.062	98.0	1.3	96.6
28	95,128	5,327	.056	108.3	1.3	107.0
29	95,002	4,750	.050	121.2	1.4	119.8
30+	94,871	4,174	.044	-	-	-

Source: United States Life Table, 1959-61, Public Health Service Publication No. 1252, Vol. 1, No. 1, Table 1, p. 8.
United States Census of Population, 1960, Final Report PC(1)-1D, Table 165, p. 371.

(8) $$T(a,x) = S(a,x)/\sum_x S(a,x)$$

It is possible, then, to treat the age-specific dropout rate, d(a), as a distribution vector, and postmultiply it to the matrix, T(a,x). The result gives a vector, d(x), which is the grade-specific dropout rate.

$$\underset{n \times m}{\begin{bmatrix} T(a_1,x_1) & T(a_1,x_2) & T(a_1,x_3) & \cdots & T(a_1,x_m) \\ T(a_2,x_1) & T(a_2,x_2) & T(a_2,x_3) & \cdots & T(a_2,x_m) \\ T(a_3,x_1) & T(a_3,x_2) & T(a_3,x_3) & \cdots & T(a_3,x_m) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ T(a_n,x_1) & T(a_n,x_2) & T(a_n,x_3) & \cdots & T(a_n,x_m) \end{bmatrix}} \times \underset{n \times 1}{\begin{bmatrix} d(a_1) \\ d(a_2) \\ d(a_3) \\ \cdot \\ \cdot \\ d(a_n) \end{bmatrix}} = \underset{m \times 1}{\begin{bmatrix} d(x_1) \\ d(x_2) \\ d(x_3) \\ \cdot \\ \cdot \\ d(x_m) \end{bmatrix}}$$

The same conversion technique can be employed to transform the age-specific mortality rate into the grade-specific mortality rate. An application of this measurement to the United States school population in 1960 is presented in Table 3. It shows the transformation matrix as well as the grade-specific dropout rate and mortality rate. The dropout rate appears to be a monotonously increasing function. It is less than 1 per thousand at the first grade, yet it increases sharply to about 300 per thousand at the end of the high school ages. On the other hand, the mortality rate is relatively constant. It is roughly about 5 per thousand, with a slight increase only after the 8th grade.

Measurement of Failure Rate

In some countries where reliable repeater figures are available the implementation of an educational planning model will involve only a very simple manipulation. For example, Liu has used the repeater statistics in Colombia to demonstrate its utility in projecting future school population (Liu, 1966). In most of the countries, however, this kind of statistics is generally lacking. So the measurement of failure rate becomes a more or less arbitrary analytical exercise. Although in a study of Australian university enrollment, Gani has suggested that an arbitrary estimation of failure rate can also show an excellent projection of future school population (Gani, 1963), it seems more commmendable to pursue a more rigorous statistical approach.

In its DYNAMOD II model, the Office of Education has proposed a less arbitrary method of estimating failure rate. The technique is very much similar to a concept which demographers call "grade retardation" (See Folger and Nam, 1967: 8-11). For instance, in the estimate of the 1st grade's failure rate, it is taken as:

(9) $$r(1) = \frac{N(8,1) + N(9,1)}{N(5,1) + N(6,1) + N(7,1) + N(8,1) + N(9,1)}$$

where N(a,x) is the school population at age a of grade x (Zabrowski and Hudman, 1967:3-4).

To what extent the measurement of retardation can represent the true picture of failure (repeater) rate remains an unanswered question, as there seems to be no empirical finding to verify such measurement so far. For this reason, we attempt to present another approach to measure the grade-specific failure rate. Although, it is not our contention that our measure is necessarily a more valid one, yet this alternative measurement may prove to be more consistent with the actual school-enrollment data.

Our approach is to make use of the model which is presented in Equation (6). The model can be changed by taking into account the measurement of separation rate as proposed in the last section. The separation rate is, by Stockwell and Nam's definition, m(x) + d(x), or mortality rate plus dropout rate. If we have the estimates of the separation rate, there can be a continuation rate, such as c(x) = 1 - (m(x) + d(x)). Then Equation (6) can be expressed as follows:

(10) $$N(x+1,t+1) = N(x,t) \cdot (c(x) - r(x)) + N(x+1,t) \cdot r(x+1)$$

It is rearranged and so it becomes:

(11) $$N(x,t) \cdot r(x) - N(x+1,t) \cdot r(x+1) = N(x,t) \cdot c(x) - N(x+1,t+1)$$

where the parameters r(x) and r(x+1) are on the left side of the equation.

Let us assume that the mortality and dropout rates derived from the census data are applicable to the rest of the decade. By doing this, it is possible to implement the continuation rate c(x) for the equation. Furthermore, the size of the student population in each grade is given by the yearly school enrollment statistics, and so the implementation of N(x,t), N(x+1,t) or N(x+1,t+1) is also possible. Consequently, the only parameters left unsolved in Equation (11) are those failure rates, r(x) and r(x+1).

It is obvious that Equation (11) is recursive with respect to every grade. Therefore, the student progression in each grade can be expressed in one of the equations; r(x) can be r(1), r(2) . . . If we substitute the coefficients N(x,t) by α, N(x+1,t) by β, and N(x,t) · c(x) - N(x+1,t+1) by λ, the following recursive equations are obtained:

$$\alpha_1 r(1) - \beta_1 r(2) = \lambda_1$$
$$\alpha_2 r(2) - \beta_2 r(3) = \lambda_2$$
$$\alpha_3 r(3) - \beta_3 r(4) = \lambda_3$$
$$\cdots\cdots\cdots\cdots$$
$$\alpha_{11} r(11) - \beta_{11} r(12) = \lambda_{11}$$

TABLE 3

DISTRIBUTION OF SCHOOL POPULATION BY AGE AND RATES OF
DROPOUT AND MORTALITY FOR EACH GRADE
UNITED STATES, 1960

Age	Grade I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII
5	.06	.01	-	-	-	-	-	-	-	-	-	-
6	.57	.06	.01	-	-	-	-	-	-	-	-	-
7	.32	.54	.06	.01	-	-	-	-	-	-	-	-
8	.04	.32	.52	.04	.01	-	-	-	-	-	-	-
9	.01	.06	.33	.51	.03	.01	-	-	-	-	-	-
10	-	.01	.06	.33	.51	.04	.01	-	-	-	-	-
11	-	-	.01	.07	.33	.51	.04	.01	-	-	-	-
12	-	-	.01	.02	.08	.33	.51	.05	.01	-	-	-
13	-	-	-	.01	.02	.08	.33	.54	.06	.01	-	-
14	-	-	-	-	.01	.02	.07	.26	.53	.06	.01	-
15	-	-	-	-	.01	.01	.03	.08	.30	.50	.06	.01
16	-	-	-	-	-	-	.01	.03	.06	.30	.53	.05
17	-	-	-	-	-	-	-	.01	.03	.07	.30	.48
18	-	-	-	-	-	-	-	.01	.01	.02	.06	.24
19	-	-	-	-	-	-	-	-	-	.01	.02	.06
20	-	-	-	-	-	-	-	-	-	-	.01	.03
21	-	-	-	-	-	-	-	-	-	-	-	.02
22	-	-	-	-	-	-	-	-	-	-	-	.01
23	-	-	-	-	-	-	-	-	-	-	-	.01
24	-	-	-	-	-	-	-	-	-	-	-	.01
25+	-	-	-	-	-	-	-	-	-	.01	.01	.07
Total	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
d(x)	.0009	.0018	.0028	.0031	.0050	.0086	.0170	.0344	.0800	.1126	.2000	.2984
m(x)	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0005	.0006	.0007	.0008	.0009

Source: United States Census of Population, 1960, Final Report PC(1)-1D, Table 168, p. 377 and computed from Table 1.

Here the failure rates r(x), where x = 1,2,3 ..., are the only parameters. An application of these equations to the United States educational system will yield eleven equations. And there are twelve unknowns in these equations. The solution is possible only if one of the unknowns is given.

It is arbitrarily assumed that the failure rate at the first grade is 0.15; to some extent such a proportion is based on the retardation concept, as DYNAMOD II suggests. With this assumption, the solution of the recursive equations yields the values of grade-specific failure rate, r(x), where x = 1, 2, 3 ... 11.

For illustration purpose, we apply this technique to the United States public school enrollment data from 1961 to 1965. Table 4 shows the estimates of failure rate at each level of education. It is found that the failure rate is relatively high in the very low grades. It declines to reach the lowest point at the 4th or 5th grade. Then it rises at the high school level.

If the quality of the data does not change much from year to year, the failure probabilities estimated through the above procedure will not show too much variation over time. Their stability contributes greatly to the reliability of the projection of future school population.

In a period from t_1 to t_n, the solution of the recursive equations give $r_t(x)$, where x = 1,2, ... 11 and r = 1,2 ... n. An examination of their mean and variance is able to indicate if a stable pattern exists. If this is the case, a trend can be fixed to these values; for instance, the trend can be a linear equation such as $r_t(x) = a + bt$. The projected value of r(x) at t_{n+1} will not differ much from the average value if the variance is not too large.

Table 4 also shows the mean, variance and the projected values of 1965-66 failure rate. Considering the fact that the quality of the data are not very satisfactory, it is interesting to note that the results do not show any particularly great deviation from year to year. As a matter of fact, the failure rates under the 7th grade are very consistent. Their variances are negligible, and after the nth grade the variance increases only slightly. Consequently, the projected failure rate, as expected, is found to be not much different from the average value of the

TABLE 4

FAILURE RATES BY GRADE IN THE UNITED STATES
PUBLIC SCHOOLS, 1961-1966

Grade	1961-62	1962-63	1963-64	1964-65	Average	Variance	Projected 1965-66
1-2	.1500	.1500	.1500	.1500	.1500	.0	.1500
2-3	.0996	.1022	.1037	.1003	.1014	.0	.1023
3-4	.0908	.0872	.0973	.0880	.0908	.0	.0912
4-5	.0856	.0792	.0927	.0810	.0845	.0	.0846
5-6	.0896	.0773	.0990	.0801	.0865	.0001	.0848
6-7	.0892	.0749	.1038	.0739	.0855	.0001	.0812
7-8	.1177	.0990	.1312	.0980	.1115	.0002	.1048
8-9	.1203	.0930	.1397	.0999	.1133	.0003	.1064
9-10	.1770	.1424	.1854	.1569	.1653	.0003	.1611
10-11	.2029	.1504	.2054	.1793	.1843	.0006	.1806
11-12	.2036	.1706	.1921	.1694	.1840	.0002	.1637

Source: Computed from Digest of Educational Statistics, 1966, Table 28, "Enrollment by Grade in Full-time Public Elementary and Secondary Day Schools," p. 24.

1961-64 period. The deviation is especially low in the elementary schools. So the result seems to indicate that there is a stable pattern of failure rates from year to year.

It appears that this technique of estimating the future failure rate may be acceptable. Therefore, the projected failure rate, along with the mortality and dropout rates, can be used in the projection of future school population. The projection procedure will be simply the application of the cohort approach as presented in Equation (6).

Summary and Discussion

This paper treats the size of student cohort as affected by three components, namely, the previous cohort's mortality, dropout, and failure. Implicitly it assumes that the educational system is a closed population. So it follows that a model of student progression can be implemented using the renewal theories in demographic analysis.

The major part of this paper attempts to assess the components of the model. It is in this aspect that this model differs from some other models of educational planning: We propose a conversion technique to obtain the grade-specific dropout and mortality rates, and we also suggest that the grade-specific failure rate can be estimated by solving the recursive equations. Through extending this type of analysis, it is possible to reach a more realistic projection of school population.

Nevertheless, this paper has not attempted to elaborate in detail all the possible approaches for reaching a successful projection of school population. For instance, one area which has been overlooked is that of stratification of the population. Males and females, whites and nonwhites, have significantly different educational parameters that change at different rates over time. Consequently, any effort to produce a useful model for other than very short-range projections must take into account the population strata.

We also conclude that a future effort shall be taken to limit the assumption that the educational system is a closed population. Conceivably this assumption may not be unrealistic in some countries where dropouts have very little chance of re-entering school; however, it is too strenuous to be applicable to the United States public school system. A study by the Bureau of Labor Statistics (1966) shows that there is a considerable number of dropouts who return to school. Therefore, we expect a future development of this paper shall be able to treat school population as an open system, and so migration and re-enrollment can be taken into account.

References

Correa, Hector. 1967. "A Survey of Mathematical Models in Educational Planning," Mathematical Models in Educational Planning, Paris: Organization for Economic Cooperation and Development.

Dressel, P. L. 1967. "Comments on the Use of Mathematical Models in Educational Planning," Mathematical Models in Educational Planning, Paris: Organization for Economic Cooperation and Development.

Duncan, Beverly. 1965. "Dropouts and the Unemployed," The Journal of Political Economy, 73:121-134.

Gani, J. 1963. "Formulae for Projecting Enrollments and Degrees Awarded in Universities," Journal of the Royal Statistical Society, Series A, 126:400-409.

Folger, J. K. and C. B. Nam. 1967. Education of the American Population, Washington, D. C.: Bureau of the Census.

Liu, B. A. 1966. Estimating Future School Enrollment in Developing Countries, N. Y.: UNESCO, Population Studies No. 40.

Moser, C. A. and P. Redfern. 1965. "A Computable Model of the Educational System in England and Wales," Proceeding of the International Institute of Statistics, Belgrade, 1965:693-700.

Organization for Economic Cooperation and Development. 1967. Methods and Statistical Needs for Educational Planning, Paris: OECD.

Pressat, Roland. 1961. L'analyse Demographique, Paris: Preses Universitaires de France.

Siegel, Jacob S. 1967. "Revised Projections of School and College Enrollment in the United States to 1985," Current Population Reports, Series P-25, No. 365.

Simon, K. A. and W. V. Grant. 1965. Digest of Educational Statistics, 1965 edition, Washington, D. C.: Office of Education.

Stockwell, E. G. and C. B. Nam. 1963. "Illustrative Tables of School Life," Journal of the American Statistical Association, 58: 1113-1124.

Stone, Richard. 1966. "A Model of the Educational System," Mathematics in the Social Sciences and Other Essays, Cambridge: The M.I.T. Press, Ch. IX, 104-117.

Thonstad, Tore. 1967. "A Mathematical Model of the Norwegian Educational System," Mathematical Models in Educational Planning, Paris: Organization for Economic Cooperation and Development.

U. S. Bureau of Census. 1966. "Projections of School and College Enrollment in the United States to 1985," Current Population Reports, P-25, No. 388.

U. S. Bureau of Labor Statistics. 1966. Out-of-School Youth-Two Years Later, Special Labor Force Report No. 71.

Wolfbein, S. L. 1949. "The Length of Working Life," Population Studies, 3:286-294.

Wurtele, Zivia S. 1967. Mathematical Models for Educational Planning, System Development Corporation, SP-3015.

Zabrowski, E. K. and J. T. Hudman. 1967. Drop-out and Retention Rate Methodology Used to Estimate First-Stage Elements of the Transition Probability Matrices for DYNAMOD II, Technical Note 28, National Center for Educational Statistics.

Zabrowski, E. K., J. T. Hudman, T. Okada and J. R. Zinter. 1967. Student-Teacher Population Growth Model: DYNAMOD II, Technical Note 34, National Center of Educational Statistics.

FINANCIAL RESOURCES OF NEGRO COLLEGE STUDENTS: SURVEY[1] DESIGN AND PRELIMINARY RESULTS

Robert F. Boruch and Alan E. Bayer

American Council on Education

During the past few years, social scientists have focused increasing attention on the black student in American colleges and universities. Pressures to do so have been generated at the governmental and institutional levels along legal social and economic dimensions. Major studies of student admissions criteria, counseling and guidance strategies and evaluation of educational standards (see Astin, 1969) are increasing. A scarcity of normative data based on a national sample of black students has been evident, however. Previous research has been informative with respect to specific environments and rather restricted samples, but limited in generalizability and in the extent to which meaningful comparisons may be made.

In order to begin study of the progress and effects of changes in educational environments on a fairly representative freshman sample, and relative to the racial issues, we have compiled information available from the ACE Higher Educational Data Bank. The data should provide base level criteria for planned examination of trends in enrollment of black students (especially through special programs), racial composition of various college types, accomplishments and persistence in the educational enterprise by black students.

In the present paper, we restrict consideration to a few specific aspects of the black and nonblack college freshman population. It is of interest to know whether students at predominantly Negro colleges differ substantially in abilities (as reflected by high school achievements) from black and nonblack students at predominantly white institutions. The information is important because competitive recruitment of the more able black students to white institutions has increased. As a result, the quality of the Negro colleges' students may decline.

Also, we should like to examine the relative financial income across institutional and racial categories. Such data can be of assistance in development of policy about financial support allocation. Finally, we would like to characterize succinctly the dependence of major financial sources of support on biographic characteristics of the students, including achievements and parental income levels.

Research Design

The normative data for this report were compiled from the third annual survey of entering college freshmen, a part of the American Council on Education's Cooperative Institutional Research Program. These data were collected by administering a four page questionnaire to more than 243,000 freshmen entering 358 institutions. The institutional stratification system for sampling(Creager, 1968) is based on the type of college, institutional affluence, and selectivity. Estimated population parameters are computed on the basis of weights derived from numbers of students within strata of the sampling design and Office of Education population counts within strata. Design specification and definitions are provided in Table 1.

Table 2 contains specification of sample racial composition and estimated distribution of population across institutional categories. Within these stratification levels, the number of black freshmen is markedly skewed. That is, one half of all colleges in the United States each enroll less than 2 percent black students among their freshmen. Eighty-eight percent have an enrollment of black students who comprise 10 percent or less of the entering class(Bayer and Boruch, 1969).

Financial Resources

Adequacy of financial aid has been a dominant theme in social science theorizing about black collegian problems. However, much of the recent information have been confined to within race and within institution descriptions. Black students are generally acknowledged to be poorer than white students, but documentation on sex differences, variation across institutional type and source of finances is rather scarce.

Consider the financial aid data for male and female students provided in Table 3. Substantial differences between racial categories within financial level are evident. Half the Negro parents' annual income level in less than $6,000. Nonwhite students acknowledge the same income bracket in only 14 percent of the population. Although racial differences across college type are not substantial, they do suggest the impact which income level has on the choice of college. Across institutional categories, the Negro colleges appear to have students with the poorest families. Predominantly white two-year colleges include the next highest fraction of low income black students, followed by predominantly white four-year colleges and the universities. Some 17 percent of black students' families have an annual income of more than $10,000, while 53 percent of the white students are in the same category.

The low parental income levels suggest, of course, that financial supplement is likely to be important in supporting the black students' education. In fact, some form of scholarship aid, grants or gifts is acknowledged to be a <u>major</u> source of income by a third of the black students. Seventeen percent of the nonblack students rely on such aid as major support. Across institutional categories, the percentages of black students with major support from scholarship funds decrease in the order that one might expect: predominantly white colleges, white universities, predominantly Negro institutions, and two-year colleges.

The increasing availability of repayable loans at minimal interest rates has been an impressive educational development during the past few years. Use of loans occurs in nearly one quarter of the population of Negro freshmen. Women receive loans somewhat more frequently in both the black and nonblack groups. The largest percentage of students who indicated that this is a major source of income occurs in the predominantly Negro institutions.[2]

Parental or family aid is a major source of support much more frequently for the nonblack students. The difference is consistent across all institutional categories except in the case of Negro institutions. Reliance on personal savings or employment is more frequently acknowledged to be a major source of funds by nonblack students.

In summary, the black-nonblack income differences are more evident than intrarace differences. Substantial numbers of black students come from families with little money. The most affluent of the black as well as the white students appear at the universities. The poorest black students enroll at the Negro institutions and (secondarily) in the junior colleges. Financial supplements rather than family aid are relied on by a majority of black students, in contrast to the nonblack. Scholarships and grants are more frequently cited by blacks as being a major source of income than any other type of aid. Non-black students rely most frequently on parental aid.

Achievements

Black students' achievement levels have been subject to much controversy and discussion. Confusion has resulted because definition of"achievements" has been confined to high school grades and grades are not entirely comparable for the black and nonblack groups. Data are presented in Table 4 and concern a variety of achievements, including grades, and relative to stratification level. The proportion of white students having A grades is about twice the fraction of black students within any institutional type. Black students in Negro colleges are not as likely to have had high secondary school grades as black students in predominantly white four-year colleges. The universities appear to be most successful in attracting larger proportions fo black students in the high grade category. Junior colleges enroll only about 2 percent of black students from this grade bracket, in contrast to the 7 to 9 percent at other types of institutions. Note that grade level comparisons are tenuous to the extent that the high schools from which students graduated differ in quality, and in predominance of a racial group. These variables are confounded with the type of institution selected by the student.

Other secondary school achievements of the freshmen are a reasonable basis for judgments of ability levels. Perhaps contrary to some expectations, a smaller percentage of black students had obtained a varsity letter in sports than the proportion of nonblack students. Larger proportions of black students had participated in National Science Foundation Summer programs, and won state or regional science contests. They were more likely to have had a major role in a play, to have received acclaim in a state or regional speech contest, or to have been elected president of a student organization. In fact, with the exception of membership in a scholastic honor society and the receipt of a varsity letter or National Merit recognition, the proportions in the various achievement categories are higher for black than nonblack students.

Differences across institutional categories are complex. Relative to black students in other institutions, higher proportions of those in the Negro colleges obtained localized recognition--winning state or regional contests; being elected to head student organizations. Within the white four-year colleges and universities, more nationally popular forms of recognition were evident, e.g. National Merit awards, varsity sports acclaim. These differences are,of course,suggestive of the admissions policies evident at most colleges: they are oriented toward rather standardized achievement modes based on the white rather than on the black student population. Systematic emphasis on multiple achievement criteria, rather than on tests alone, is warranted insofar as equitable distribution of students among colleges is an objective of the current government and educational effort.

Regression Analyses

We have used stepwise regression analyses in order to assay the extent to which the mode of students' major financial aid is dependent on other observable factors. Immediate requirements are largely descriptive and very simplistic: specification of the nature of the relation, and of the moderating effect of stratification factors on functional relationships.

An intermediate objective is to assess the extent to which regression(commonly used in the educational literature) is affected by measurement error. Lacking estimates of reliablility for each variable, a cross validation technique (Wolins, 1967) is used to strengthen the credibility of the computed multiple correlation. An original (validation) sample comprises the basis for specification of predictor variables in the linear function. A second (cross validation) sample and the previously specified predictor variables, are used to compute an unbiased $\hat{R}\, y\hat{y}$, conditional on the second example (Wherry, 1931).

Four dependent variables, and not mutually exclusive, are considered: personal savings, parental aid, loans and scholarships (Appendix I). Each is scored on the basis of student acknowledgement that the particular income is a major source of support, a minor source, or not a source of support at all. In addition to parental income and student achievements, other independent (dummy) variables have been introduced: parental educational levels, aspirations and past behavior. Samples (approximately 1,000 students each) were systematically drawn for the categories of Negro men and women. The samples were further categorized by institutional type.

Table 5 contains some results of the regression analyses; independent variables are presented in rank order of standard regression weight size. Generally, the percentage sums of squares accounted for in the cross validation sample is larger for men than for women. Decrements in the estimated population parameter,[3] from validation to cross validation, suggest somewhat higher measurement error in the data on women. The absolute magnitudes of the $\hat{R}^2 y\hat{y}$ are rather unimpressive, but they are a bit higher than results typically obtained in this type of research.

The differences between the regression

equations across sex and institutional categories implies, of course, that stratification attributes ought to be recognized in such analyses. The variation of the $\hat{R}^2 y\hat{y}$ with in sex and type college is fairly stable across all categories considered.

Consider now the predictor variables with higher regression coefficients. Substantial reliance on personal savings (variable 15) is associated with older students for both sexes and within universities and Negro colleges. The choice of colleges, and for black men in predominantly white universities. Achievement variables enter negatively in most cases so one can infer that little or no reliance on personal savings is associated with academic achievements of various types.

Students' dependence on parental aid (16) is a function of age and parental income for all institutions and sexes: younger students from more affluent families acknowledge this type of support. The mother's education appears to be associated with men's acknowledgement of this source of support, and father's education is a determinant for women, at the Negro colleges and universities. Achievement variables are weighted negatively, suggesting that lower achievers of the type of award considered here are relying more heavily on their parent's income.

The extent to which we can explain loan usage (17) as a function of other variables used here is markedly limited. The relatively low R yy may be explicable in terms of the nonuniformity of banking practices and in the extent to which students investigate this source of income. Loan usage is associated with lower family incomes. Lower costs of college are a determinant of the responses for men and women within the universities, but not for the other categories. Younger students appear to be relying on loans.

Student acknowledgement of major dependence on scholarship aid (18) is somewhat more predictable than either reliance on loans or on personal savings. Within the Negro colleges, parental aid is weighted negatively for both men and women, suggesting that poorer students are receiving such support. For men, the variable is a function of mother's education and high school grades; athletic scholarships are probably substantial since winning a sports letter and choosing a college on the basis of athletic program are predictor variables. Important independent variables within the women's category include high school rank and the affirmation that choice of college was based on it's academic reputation. For men at the universities, major reliance on grants is largely a function of better high school grades and awards, and coming from low income families in which mother's education is low. For university women, the important predictors are family income and various achievement and study habit variables --- art or science recognition, National Merit recognition. Within the white 4 year colleges, men and women relying on scholarship aid come from families with smaller incomes. Achievement predictors include membership in an honorary society and art recognition (men), and awards in regional science contest, high school rank, having a major part in a high school play.

These findings represent a caricature of the financial and achievement attributes of the of the Negro freshmen population. Forms of financial support are predicatable in varying degrees from biographical data. The major reliance on scholarship aid and parental support is associated with plausible and well defined student attributes but the extent to which we can predict is not impressive. Further scrutiny of these data will clarify the meaning of the associations among variables, and may provide information for enchancing our ability to predict the dependent variables.

Footnotes

1. This paper was supported, in part, by National Science Foundation Grant, GR-57.

2. During the summer of 1969, the popular press frequently reported that banks and loan companies are curtailing or eliminating student loan programs. It appears that black college students will be adversely affected by these measures, unless other sources of income are made available.

3. $\hat{R}^2 y\hat{y} = 1 - (1-\hat{R}^2 y\hat{y}) \dfrac{(N - 1)}{(N - m - 1)}$

References

1. Astin, A.W. Racial considerations in college admissions. The Campus and the Racial Crises, Washington: American Council on Education, 1969 (in press).
2. Bayer, A.E. and Boruch, R.F. Black Student in American Colleges. ACE Research Reports. Vol. 4, No. 2, 1969. Washington, D.C.: American Council on Education, 1968.
3. Creager, J.A. General purpose sampling in the domain of higher education. ACE Research Reports, Vol.3, No. 2, 1968.
4. Wherry, R.J. A new formula for predicting the shrinkage of multiple correlation. Annals of Mathematical Statistics. 1931, 2, 440-451.
5. Wolins, Leroy. The use of multiple regression procedures when the prediction variables are psychological tests. Educational and Psychological Measurements, 1967, 27, 821-828.

Table 1: 1968 ACE Sample

Stratification Cell[a]	Institutions	
	Popu-lation	Used in Norms
Univeristies		
Selectivity:		
Less than 500	30	10
500-549	39	15
550-599	45	15
600 or more	50	25
Unknown	130	11
Public Colleges (4 year)		
Selectivity:		
Less than 450 and Unknown	154	10
450-499	67	9
500 or more	73	14
Private Non-sectarian (4 year)		
Selectivity:		
Less than 500 and unknown	197	24
500-574	44	7
575-649	54	18
650 or more	48	27
Roman Catholic (4 year)		
Selectivity:		
Less than 500 and unknown	111	15
500-574	75	13
575 or more	42	15
Protestant (4 year)		
Selectivity:		
Less than 450 and unknown	119	14
450-499	54	7
500-574	68	13
575 or more	48	14
Two Year Colleges		
Selectivity less than 400	87	4
Selectivity 400-499	63	11
Selectivity 450 or more	57	8
Expenditures /less than $1000	192	12
Expenditures /$1000-$1249	39	4
Expenditures /$1250 or more	52	7
Selectivity or Expenditures unknown	272	17
Negro Colleges		
Public	38	7
Private	55	12

[a]In addition to the specifications by the U.S. Office of Education of level and type of control, the stratification design includes institutional per-student expenditures (for students who completed the National Merit Scholarship Qualifying Test).

Table 2: Number of Institutions and Students Used in Computing Norms

Norm Group	Number of Institutions in Norms	Number of Students in Norms		Weighted Number of Students		Percent of Weighted Total Who are Men	
		Black	Nonblack	Black	Nonbkack	Black	Nonblack
All Institutions	358	12,300	230,582	84,058	1,386,369	46.0	57.2
Two-Year Colleges	63	1,535	34,342	19,093	426,337	55.4	61.2
White Four-Year Colleges	200	1,996	75,820	15,373	536,680	45.1	53.0
Negro Four-Year Colleges	19	5,384	349	36,071	2,560	42.5	69.0
Universities	76	3,385	120,071	13,521	420,792	42.9	58.4

Table 3: Weighted Percentages For Entering College Freshmen

	Predominantly White Two-year Colleges		Predominantly White Four-year Colleges		Negro Four-year Colleges		Predominantly White Universities	
	Black	Nonblack	Black	Nonblack	Black	Nonblack	Black	Nonblack
MEN								
ESTIMATED PARENTAL INCOME								
LESS THAN $4,000	25.9	6.1	25.2	4.2	37.2		19.0	3.5
$4,000 - $5,999	30.8	12.6	24.6	8.9	25.1		21.8	7.0
$6,000 - $7,999	18.2	19.6	18.9	15.3	14.2		21.4	12.5
$8,000 - $9,999	10.9	19.3	10.1	18.2	9.2		14.2	16.3
$10,000 - $14,999	9.8	26.7	13.2	29.0	8.8		15.3	29.9
$15,000 - $19,999	3.7	8.7	4.0	11.6	3.4		5.2	13.2
$20,000 - $24,999	0.3	3.4	1.6	5.1	1.0		1.3	6.8
$25,000 - $29,999	0.1	1.3	1.1	2.7	0.4		0.6	3.4
$30,000 OR MORE	0.3	2.3	1.2	5.0	0.7		1.2	7.4
MAJOR SOURCES OF SUPPORT								
PERSONAL SVGS OR EMPLOYMENT	32.2	46.1	17.5	30.1	18.8		19.3	30.4
PARENTAL OR FAMILY AID	21.9	37.1	21.3	46.9	28.0		33.3	55.7
REPAYABLE LOAN	15.0	7.9	28.6	16.4	29.0		16.5	10.5
SCHOLARSHIP/GRANT/OR OTHER GIFT	32.8	10.9	49.8	22.2	35.1		42.5	17.9
WOMEN								
ESTIMATED PARENTAL INCOME								
LESS THAN $4,000	29.4	7.4	27.8	4.7	38.0		22.1	3.3
$4,000 - $5,999	26.3	13.1	19.9	8.5	24.2		24.5	7.0
$6,000 - $7,999	19.9	18.7	18.5	14.5	14.1		20.2	11.5
$8,000 - $9,999	9.8	18.1	12.1	16.2	9.2		12.8	14.7
$10,000 - $14,999	9.5	24.9	14.3	28.1	8.4		13.2	29.7
$15,000 - $19,999	2.9	9.8	4.8	12.7	3.3		4.5	14.7
$20,000 - $24,999	1.7	3.7	1.3	6.3	1.9		1.5	8.2
$25,000 - $29,999	0.0	1.8	0.4	3.2	0.6		0.7	3.9
$30,000 OR MORE	0.5	2.7	0.9	5.8	0.4		0.5	6.9
MAJOR SOURCES OF SUPPORT								
PERSONAL SVGS OR EMPLOYMENT	30.4	27.7	12.7	15.3	10.7		13.3	15.4
PARENTAL OR FAMILY AID	29.6	55.8	31.7	62.1	34.5		43.7	69.9
REPAYABLE LOAN	13.0	10.0	27.6	18.8	37.4		18.2	11.6
SCHOLARSHIP/GRANT/OR OTHER GIFT	24.0	11.2	54.4	21.5	29.4		37.4	16.9

Table 4: Weighted Percentages For Entering College Freshmen

MEN	Predominantly White Two-year Colleges		Predominantly White Four-year Colleges		Negro Colleges	Predominantly White Universities	
	Black	Nonblack	Black	Nonblack	Black	Black	Nonblack
AVERAGE GRADE IN GHIGH SCHOOL							
A OR A+	0.3	0.4	0.6	3.9	1.0	1.9	6.3
A-	0.8	1.3	3.2	7.7	3.0	5.5	10.9
B+	6.7	4.7	10.7	15.2	11.1	13.9	18.1
B	11.7	14.4	18.4	22.6	19.2	21.3	23.5
B-	11.0	15.3	16.7	17.7	16.3	19.0	16.5
C+	27.4	26.4	26.3	18.1	27.1	23.8	14.4
C	38.9	34.7	22.1	14.0	21.4	14.0	9.8
D	3.4	2.8	2.0	0.8	0.9	0.6	0.6
SECONDARY SCHOOL ACHIEVEMENTS							
ELECTED PRESIDENT STDT	14.7	12.1	27.6	22.7	32.4	28.0	24.5
HIGH RATING STATE MUSIC	6.7	6.2	9.9	8.5	12.1	11.1	9.9
STATE/REGIONAL SPEECH CONTEST	1.7	2.9	5.7	5.4	7.8	6.8	6.7
MAJOR PART IN A PLAY	15.5	12.6	19.7	17.6	34.5	19.4	16.9
VARSITY LETTER (SPORTS)	53.3	40.5	58.3	50.8	42.0	45.0	44.2
AWARD IN ART COMPETITION	7.2	4.9	7.9	4.2	7.8	7.3	4.1
EDITED SCHOOL PAPER	3.4	4.8	9.1	9.1	11.2	10.3	10.0
HAD ORIGINAL WRITING PUBLISHED	5.7	7.1	18.3	14.7	9.7	14.7	16.3
NSF SUMMER PROGRAM	1.4	0.4	1.7	1.0	2.3	2.1	1.6
ST/REGIONAL SCIENCE CONTEST	1.6	1.4	3.1	2.5	6.7	5.0	3.5
SCHOLASTIC HONOR SOCIETY	3.9	4.4	15.0	23.0	17.1	22.7	31.0
NATIONAL MERIT RECOGNITION	2.6	1.6	8.6	8.1	5.1	10.6	12.3

Table 4: Weighted Percentages For Entering College Freshmen (cont'd.)

WOMEN	Predominantly White Two-year Colleges		Predominantly White Four-year Colleges		Negro Colleges	Predominantly White Universities	
	Black	Nonblack	Black	Nonblack	Black	Black	Nonblack
AVERAGE GRADE IN HIGH SCHOOL							
A OR A+	0.8	1.9	2.5	7.2	2.0	2.5	10.1
A-	2.7	4.3	7.1	13.8	7.0	7.9	16.7
B+	11.3	11.9	17.6	23.7	20.9	18.0	24.0
B	21.8	26.9	23.5	28.5	25.4	25.9	25.6
B-	17.2	17.9	18.9	13.0	15.5	16.1	11.4
C+	22.5	19.1	18.4	9.1	17.6	17.8	7.9
C	22.5	17.4	11.8	4.6	11.2	11.6	4.3
D	1.3	0.5	0.2	0.1	0.4	0.2	0.1
SECONDARY SCHOOL ACHIEVEMENTS							
ELECTED PRESIDENT STDT ORGNZ	12.7	13.7	24.9	22.5	31.8	23.4	23.2
HIGH RATING STATE MUSIC	6.0	8.7	10.7	13.0	12.0	10.4	14.7
STATE/REGIONAL SPEECH CONTEST	6.9	4.3	7.6	6.6	8.9	6.8	7.7
MAJOR PART IN A PLAY	17.0	13.1	20.6	18.5	34.3	17.5	18.4
VARSITY LETTER (SPORTS)	13.4	13.1	12.0	14.1	9.8	9.5	12.7
AWARD IN ART COMPETITION	4.2	5.7	6.0	6.6	4.4	4.8	7.1
EDITED SCHOOL PAPER	8.1	10.1	15.7	17.3	16.1	13.1	17.7
HAD ORIGINAL WRITING PUBLISHED	10.2	13.7	20.2	22.3	14.0	18.5	22.9
NSF SUMMER PORGRAM	0.4	0.2	2.3	0.6	1.5	1.3	0.8
ST/REGIONAL SCIENCE CONTEST	2.4	1.2	5.4	2.1	7.2	3.0	2.6
SCHOLASTIC HONOR SOCIETY	11.5	13.7	33.1	38.8	31.9	30.1	45.1
NATIONAL MERIT RECOGNITION	2.9	2.7	11.7	8.5	6.5	11.1	10.8

Table 5

College Type/Sex	y	Rv	Rcv	$\hat{R}y\hat{y}$	d.f.	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}
Negro	15	.34	.31	.10	418	1	-37	-29	38	36	13	-24	-40	8	-4				
4-year/	16	.52	.43	.17	419	21	20	-1	-23	12	-48	-25	5	38					
Negro	17	.34	.31	.10	420·	-21	-7	-1	-20	-32	40	-26	-13						
Men	18	.42	.41	.15	418·	-20	2	37	-38	50	-21	7	43	-24	32				
Negro	15	.24	.21	.03	595	1	12	48	25	38	3	-27	-8						
4-year/	16	.45	.41	.15	592	21	19	-3	-43	-12	-24	-34	-1	-22	25	49			
Negro	17	.40	.27	.06	592	-21	34	-20	-14	26	-4	37	30	-32	-1				
Women	18	.42	.37	.12	591	23	39	-21	48	3	36	37	33	29	22	2	5		
White	15	.49	.26	.04	299	38	-7	40	1	-6	-26	14	47	-46					
Univer/	16	.58	.49	.22	299	21	-1	29	20	-46	47	33	-44	52					
Negro	17	.42	.21	.02	298·	-46	47	4	-1	-38	-21	35	6	36	54				
Men	18	.55	.46	.18	296·	-21	-38	2	-20	-6	53	-29	13	52	51	39	-47		
White	15	.32	.20	.02	356·	-25	-11	1	21	34	-36	27							
Univer/	16	.54	.43	.17	336	21	-31	-1	-13	-8	-14	44							
Negro	17	.38	.24	.02	331·	-38	46	-21	23	-27	6	-13	-7	54	53	29	-10		
Women	18	.48	.35	.10	332·	-21	13	-44	32	5	-46	-26	34	14	8	-12			
White	15	.49	.35	.08	166	-3	48	8	-4	-54	-10	-24	33						
4-year/	16	.68	.38	.10	167	21	19	-22	-41	34	-7	-35							
Negro	17	.47	.22	.01	167·	-51	49	-21	-1	-3	-25	34							
Men	18	.54	.49	.19	163·	-21	13	-36	52	-48	34	-8	-19	9	26	49			
White	15	.49	.21	.00	195	38	-43	54	-46	44	-26	-11	27	-21	-24				
4-year/	16	.66	.48	.19	194	21	-4	-8	-1	-31	-11	-38	48	-5	27	-53			
Negro	17	.49	.36	.08	193·	-21	48	8	35	-42	45	-12	46	-25	34	1	-38		
Women	18	.57	.44	.14	192·	-21	23	31	-12	11	-6	35	-41	40	34	36	-48	-24	

Appendix I
Variables Included in Regression*

1 Age (Lo : Hi :: 1:9)
2 High school grades (Lo : Hi :: 1:8)
3 Elected president stdnt organ
4 Hi rating music contest
5 Part in state speech contest
6 Major part in play
7 Won varsity letter
8 Won art award
9 Edited school paper
10 Published poem, essay
11 NSF summer program
12 Placed in state science contest
13 Member of honorary society
14 Won certificate of merit (NMSQT)
15 Personal savings
16 Parental aid
17 Repayable loan
18 Grant/gift
19 Father's education (Lo : Hi :: 1:6)
20 Mother's education (Lo : Hi :: 1:6)
21 Parent's income (Lo : Hi :: 1:9)
22 Religious preference
23 Academic rank in graduate class
24 Outlined reading assignment
25 Memorized without understanding
26 Shared notes with students
27 Extra credit work
28 Tests for practice
29 Played musical instrument
30 Anti-war protest
31 Racial protest
32 Administrative policy protest
33 Extra reading for course
34 Teacher recommended college
35 Graduate rep recommended college
36 Professional counselling recommended college
37 Athletic program reason for college choice
38 Low cost reason for college choice
39 Academic reputation reason for college

(Aspirations)

40 Become accomplished in performing arts
41 Become authority on specified subjects
42 Attain recognition for contributions
43 Become expert musician
44 Become expert in finance
45 Have admin. responsibility
46 To be well off financially
47 Aspire to help others
48 To be in organization such as Peace Corps
49 Become outstanding athlete
50 Become community leader
51 Make theoret. contribution
52 To write original work
53 To create artistic work
54 To be success in business of own

* Dichotomous variables scored 2 (yes) or 1 (no). Variables 15, 16, 17, 18 scored 3 (major dependence), 2 (minor dependence), 1 (not at all) on sources of support indicated.

AN EMPIRICAL GROUPING OF JOB FAMILIES

Marion F. Shaycoft, American Institutes for Research[1]

Project TALENT is a long-term educational research project that started about ten years ago and is expected to continue about 25 years altogether. The project has now reached the point where questions of whether jobs can be grouped into families, and if so how, are important.

Project TALENT: The source of the data

In 1960, a comprehensive battery of tests and questionnaires lasting two full days was administered to about 400,000 students in over 1000 secondary schools. This very large sample, consisting of the students in grades 9, 10, 11, and 12 in a stratified random sample of all secondary schools -- public, parochial, and private -- in the United States, has been followed up by questionnaire one year and five years after high school graduation. The plans of the project include further follow-ups ten and twenty years after high school. Though the follow-up questionnaires cover a fairly wide range of areas, they focus most sharply on post-high school education and on jobs and long-range career plans.

One of the major purposes of the study is to provide a basis for improving vocational and educational guidance in the high schools by finding out what kinds of aptitudes, interests, achievement levels, personality traits, and other characteristics manifest at the high school stage of development are predictive of success in specific occupations. In the questionnaires the basic questions to elicit information about career plans were:

(a) What occupation do you plan to make your life work? Be as specific as possible. For instance, if military service specify type of work. ____________________

____

____

(b) What steps have you taken in this direction? (Mark as many as apply.)
- a. I now have or have had a regular job in this field.
- b. I now have or have had a job as a trainee in this field.
- c. My present job may lead to work in this field.
- d. I am doing or have done volunteer work in this field.
- e. I have had special training or education in this field.
- f. None of the above.

(c) If you have had special training or education in this field, how or where did you get it? (Mark as many as apply.)
- a. In high school.
- b. In college as an undergraduate.
- c. In graduate school or professional school after college.
- d. In some other kind of school, since high school.
- e. An apprenticeship program.
- f. On-the-job training (informal or formal).
- g. An informal program: reading or other independent study.
- h. Some other way.
- i. I have had no special training or education in this field.

Project TALENT's procedures and instruments are documented elsewhere [1,2,3,4,6,7].

Purpose of grouping

The career fields indicated by Project TALENT participants when they were followed up five years after high school were initially coded into nearly a thousand categories in order to retain as much information as possible, and to permit subsequent collapsing of categories in a multiplicity of ways. It was recognized, of course, that some collapsing would be necessary, since 1000 categories would be far more than could be handled conveniently in any data analysis. Although we could have bypassed the thousand-category stage altogether, this would have been undesirable because having more detail initially than would be needed in any one analysis would allow maximum flexibility in combining categories later on, and thus would permit different kinds and degrees of condensation of categories for different purposes.

As the first step in reducing the number of categories for potential use in educational and career guidance of high school students, the original categories were collapsed on a judgmental basis to about 250. It was felt that this was the most that could be done safely on the basis of subjective judgment; that any further combining of groups should be based on empirical data. More specifically it was hoped that on the basis of empirical data the 250 categories could be condensed into a much smaller number of groups, such that the categories combined in a single group would be relatively homogeneous in terms of the patterns of aptitudes, abilities, achievement levels, interests, personality traits, and background factors characteristic of their members. What was sought was relative homogeneity within groups and heterogeneity among groups with respect to scores on 64 cognitive variables and 45 noncognitive variables from the TALENT battery. This might simplify educational and vocational guidance to some extent, by making it possible for guidance counselors in high schools to advise the student in terms of families of jobs for which he is suited, rather than in terms of a small or large number of specific career fields from which to choose.

Besides convenience, there were other reasons for wanting to determine some "job families." Despite the very large number of cases Project TALENT had started with initially, some of the less usual career fields had very few cases in them, and therefore might not provide stable data unless they were combined with other closely related groups.

Methodological consideration in using hierarchical analysis

Hierarchical analysis seemed like a promising way of establishing job families. But the term hierarchical analysis doesn't cover just one specific procedure. Rather it represents a whole family of procedures -- so that once one has decided to use hierarchical analysis his decisions have just begun.

In a hierarchical grouping procedure, one generally starts with a matrix showing the degree of difference (or similarity) between each individual and each other individual. The hierarchical analysis procedure operates in a stepwise fashion, to combine individuals (or groups) for whom the index of difference is as low as possible, or the index of similarity as high as possible. Difference may be expressed in a number of ways: as distance between individuals (or groups), as dispersion of the combined group, or perhaps as a ratio of distance to dispersion. Likewise similarity can be expressed in numerous ways -- as amount of overlap between groups, for instance, or perhaps as correlation between group means where the means are expressed as ipsative scores.

Some of the methodological considerations in hierarchical analysis are discussed below.

1. Deciding on the order of merging

   To get an idea of some of the problems, and the point of view underlying the methodological decisions that were made, let's look at Figure 1. (In the interest of simplicity this diagram and all others in this paper are limited to the two-dimensional or two-variable situation, but the problems and conclusions are readily generalizable to any number of variables.) Each circle (or ellipse) represents a group -- let's say a group of people in the same career field. The radius of the circle represents the dispersion of the group, and the dimensions of the ellipses have an analogous interpretation, so that each of the circles and ellipses in the diagram encloses the same proportion of cases in the group it represents. Let's suppose, for instance, that each circle or ellipse encloses 95 percent of its group. Which pair of the groups in Figure 1 should be combined first?

   a. Should groups A2 and B2 be combined before groups A1 and B1 are? The centroids of A2 and B2 are closer together than those of A1 and B1, and it also appears that the A2-B2 combination would be more compact than the A1-B1 combination. But in terms of degree of overlap the A1-B1 combination seems about as good as A2-B2.

   b. Now let's look at A3 and B3, the two very compact distributions represented by the tiny circles toward the bottom of the chart. Their centroids are about the same distance apart as the centroids of A1 and B1. But probably in view of the differential overlap rates, A1 and B1 are better candidates for merging than are A3 and B3.

   c. Should Groups A4 and B4 be merged before A2 and B2 are, or should the A2-B2 merging take precedence? The A4-B4 pair of circles looks about the same (in size and overlap) as the A2-B2 pair but the A4 and B4 groups contain 200 and 400 cases respectively while the A2 and B2 groups are much smaller, containing only 50 cases each. Does this affect their mergeability? Remember that in the diagram the radius of a circle represents the dispersion of the entire group rather than sampling error. Because the A4 and B4 centroids have smaller sampling errors than the A2 and B2 centroids, the distance between centroids $\bar{A}4$ and $\bar{B}4$ is statistically significant to a greater degree than the distance between $\bar{A}2$ and $\bar{B}2$. But this wouldn't be any reason for merging the smaller A2 and B2 ahead of A4 and B4. We aren't trying to limit the merging to groups that do not differ significantly. We are quite willing to admit that probably no two of the populations represented by the various groups are identical in their statistical characteristics. In other words it is quite likely that all the groups -- even the ones we merge -- differ significantly. The important questions is not whether they differ, but how much they differ, since we would like the merging confined to groups whose differences are relatively small. The A2-B2 pair and the A4-B4 pair are equally good candidates for merging.

   d. What about the mergeability of the groups represented by the A5-B5 pair of ellipses in comparison with the mergeability of the A6 and B6 groups, represented by ellipses of similar size, shape, and orientation? Note that the distance between centroids $\bar{A}5$ and $\bar{B}5$ is the same as that between centroids $\bar{A}6$ and $\bar{B}6$. But despite all these similarities between the A5 and B5 pair and the A6-B6 pair there is vast difference in their mergeability. Groups A6 and B6 overlap substantially while groups A5 and B5 hardly overlap at all. This corresponds to the fact that in the case of the A6-B6 pair the dimension in which the distance between centroids lies is the dimension in which within group dispersion is largest, while in the case of the A5-B5 groups the opposite is true. There might be some justification, then, for merging A6 and B6, but there is probably none for merging A5 and B5.

e. How about A7 and B7? The distance between centroids is about twice as large for this pair as the A6 and B6, but the dispersions are also about twice as large, and the A7-B7 configuration is entirely proportional to the A6-B6 configuration. The 90° rotation, reversing the relationship to the horizontal and vertical dimensions, of course doesn't alter this. Thus the two pairs are equally mergeable.

f. As for A8 and B8, this pair is about the same general configuration, except for the 45° difference in angular orientation and the greater overlap, as the A5-B5 pair. Actually the A8-B8 pair has the same amount of overlap as A6-B6 or A7-B7, and is equally mergeable.

To recapitulate our conclusions in regard to Figure 1:

(1) A1-B1, A2-B2, and A4-B4 are all equally mergeable.

(2) A6-B6, A7-B7, and A8-B8 are all equally mergeable, and each of these pairs is far more mergeable than A5-B5.

(3) A3 and B3 should not be merged.

If these are the decisions we want our hierarchical procedure to result in, what kind of formula should be used as the basis on which merge decisions are made? Mere distance between centroids, merging the two groups whose centroids are closest together geometrically, won't give the desired result. Nor will minimizing any kind of variance measure such as the mean square distance of points in the new combined group from the centroid of the new group.

Formula 8 in the Appendix[2] gives the geometric distance between two centroids, plotted in n-dimensional space. This is the generalized Pythagorean Theorem. Formula 9 gives the distance between centroids when each dimension is appropriately scaled in terms of the standard deviation of the distance between centroids along the dimension. The formula 9 value (or any monotonic transformation of it such as its square, given by formula 7) <u>will</u> give the desired results. Formula 7, therefore, is the one we would have liked to use in our research on grouping of jobs. Because of practical considerations, however, we actually had to do this research as a two-stage operation, using formula 5, which gives the square of the distance between centroids, in the preliminary stage and formula 7 in the final stage. The preliminary stage consisted in hierarchical analysis; the final stage consisted in using formula 7 to check on the tentative groupings from the hierarchical analysis and making modifications where appropriate.[3]

Practical considerations, such as limitations on computer capacity, precluded use of formula 7 in the hierarchical analysis itself, since we wanted to analyze up to 173 job groups in terms of as many as 64 variables. To do this with any kind of reasonable efficiency would have required about 2½ times as big a computer as formula 5, and about 2½ times as big a computer as we had available. Although formula 5 was known in advance not to be the ideal formula for our analysis, it did turn out to be a very useful one and to work well. Several other formulas (formulas 11-14) were tried out as alternatives to formula 5. All these alternative formulas[4] represented efforts to achieve partially the advantages of formula 7 over formula 5 without requiring a computer with any more core capacity than formula 5 requires. However most of these alternative formulas, when applied to our data, turned out to give roughly the same results as formula 5, and <u>none gave any better results</u>.[5] In the interests of simplicity, therefore, formula 5, which was by far the simplest of the formulas tried in the hierarchical analysis program, was the one used operationally.

Therefore the input to the hierarchical analysis was a matrix of d^2 values. At each stage of the hierarchical analysis the two jobs or job groups were combined whose centroids were closest together. Formula 6 was used to compute the square of the distance between the new group thus formed and each of the other groups.

2. <u>Kind of scores scales to be used</u>

Having decided what formulas to use (formulas 5 and 7) to express difference between groups, the next question is what kind of scores to apply the formula to. In other words should we use the initial test scores in their raw form? Or should they be converted to some kind of factor score, or discriminant function, or some other kind of derived scores? And should the number of variables be reduced through some such procedure as converting to factor scores and then using only the first few factors? It was decided that orthogonal scores would be quite necessary but that dimension reduction would be extremely undesirable. The advantage of orthogonal variables was that they would result in a meaningful d^2 matrix uncontaminated by the effects of correlation.

But what <u>kind</u> of orthogonal variables? It was decided that our needs in this direction would be served best by principal components, scaled in the usual way -- with zero means and unit standard deviations. The possibility of using discriminant functions instead of principal components in order to get orthogonal variables was given careful consideration and rejected. Discriminant functions, unlike principal components, are normally scaled in such a way that their variances are proportional to their overall effectiveness in discriminating among groups. Principal components with uniform standard deviations of 1 obviously lack this feature, as far as individuals are concerned. But for group centroids this deficiency is self-correcting, since the dispersion among groups means is of course far greater for the principal components that discriminate effectively among groups than for those that don't.

3. Should a scale-free method be used?

All the problems concerning choice of a measure of geometric distance and/or dispersion suggested an entirely different possibility -- the possibility that perhaps we should bypass all these considerations of a strictly metric nature by using a method that is both simple and computationally invariant under monotonic transformation of the data -- Johnson's ultrametric maximum method, for instance, or his ultrametric minimum method [5]. After careful consideration it was decided that these methods were not suitable for the kind of data we had. Let's look at some strictly hypothetical data illustrating one of the disadvantages. Figure 2 shows an example of the ultrametric maximum method and Figure 3 the ultrametric minimum method. Both are artificial data, but the results are bizarre enough to give some idea of the sorts of peculiarities that may result when useful parametric data are ignored. According to the maximum method, the pair of groups to be merged is the pair for which the maximum distance between a point in one group and a point in the other group is smallest. This procedure is intended to yield maximally compact groups -- but that isn't always the actual result. Figure 2a presents a 14-point grouping problem. Three of the 14 points of Figure 2a form a very compact cluster at the left, while the other 11 form a somewhat more diffuse cluster at the right. Two solutions are presented -- one in Figure 2c and the other in Figure 2d. Figure 2b shows two intermediate stages of grouping which would occur if the d^2 criterion (square of geometric distance) were used. (The first of the two would also occur with the ultrametric maximum method.) Figure 2c shows the final grouping resulting from the maximum method. Three of the points that seem rightfully to belong in the right-hand cluster are joined with the left-hand cluster. This strange result seems to be a blatant instance of "empire-building" by group J. Figure 2d shows the more normal results obtained through the use of d^2. (Table 1 is the distance matrix corresponding to Figure 2. It shows the distances themselves, not their squares, but this makes no difference in the results.)

The ultrametric minimum method merges those two groups closest to each other, when the closeness of two groups is defined as the distance between the two points that are closest to each other. Figure 3 shows a dumbbell configuration of points with a small hexagonal arrangement near one end of the dumbbell. As shown at the bottom of Figure 3, when the minimum method is used the entire dumbbell turns out to be one cluster, even though the two obvious clusters of the dumbbell are actually connected by only the most tenuous chain of points. If any single point on the chain joining the two ends of the dumbbell were dropped from the configuration the dumbbell would collapse into two parts immediately.

4. When to stop merging

Getting back to our old-fashioned metric data of formulas 5 and 7, how does one decide when these values have become too large to warrant further combining of groups? For formula 7 the answer lies in the fact that there is a way of interpreting the numerical values in geometric terms. This is shown in Figure 4. As in Figure 1, the centers of the Figure 4 circles represent the centroids and the radii are assumed to indicate the dispersion.

Since the circles include almost everyone in the group, the last pair corresponds to two groups that have almost no overlap. The D^2 is 8 (where D^2 is defined by formula 7). The first pair, which has very substantial overlap, has a D^2 of only .32. It seems undesirable to combine jobs that have a D^2 much above 1.50, because the people in them are too different to be lumped together in one heading. It was therefore decided to apply this rather stringent criterion to our empirical data, in determining what career groups to merge. There didn't seem to be any compelling reason for forcing every job to be combined in a "family" with other jobs if there were some that didn't fall into natural clusters.

So much for the methodological decisions on grouping. Now let's get to our actual empirical study, applying the methods we decided were most suitable for the grouping of jobs.

General procedures in the empirical study

The grouping study was based on the test scores and other data collected on 14123 grade 12 boys who responded to the follow-up questionnaire sent them five years after high school. To get orthogonal variables for the total group, two principal components analyses were carried out -- one for the cognitive variables and one for the noncognitive. As many principal components were obtained as there were variables in the battery -- 64 for the analysis of cognitive variables and 45 for the noncognitive. All of these are being used, since there seems to be no advantage to dimension reduction in this situation and rather substantial disadvantages in terms of the potential loss of information that could result from reducing 64 or 45 variables to a substantially smaller number. Preliminary results are presented in this report.

After the respondents to the follow-up questionnaire were classified into the a priori categories on the basis of their career plans, the groups were "purified" by eliminating cases where the alleged career plan seemed to have little basis in reality. For instance anyone who indicated five years after high school that he intended to become a physician was excluded from the purified group of prospective physicians if he had not even entered college yet. Objective criteria were set up in advance for each career category, defining what kinds of responses, if any, would result in exclusion from the purified group.

Tentative job families were established on the basis of hierarchical analysis of the cognitive data and eliminated or modified if the analysis of the noncognitive data didn't confirm them. (Actually as things turned out, the cognitive and noncognitive results agreed very well. Hardly any groups had to be eliminated or changed on the basis of different findings from the two analyses.)

In establishing tentative groups on the basis of results of the hierarchical analyses, some liberties had to be taken with the hierarchical model, because the data appeared not to cluster in groups which fitted this model very closely. There appeared to be only a very small number of nuclei of clusters and before these nuclei acquired a large number of "satellite groups" in the hierarchical development they tended to coalesce. Thus, depending on where the merging process was stopped we were presented with the choice of either a small number of clusters, most of them including only about three or four groups each, supplemented by a very large number of separate groups (i.e. single-group "clusters") or, alternatively, one very large cluster, which has swallowed up the few smaller separate clusters and has also swallowed up most of the separated groups. If we were to hold strictly to the hierarchical model there didn't seem to be any happy medium between these two extremes. However study of the hierarchical data led to the conclusion that meaningful and useful clusters could be established from the long chain of careers groups that the hierarchical analysis tended after a while to yield, by breaking the chain at carefully chosen points to split it into several sections. In a few instances there was some ambiguity as to the exact point at which it would be best to break the chain, because the career group in the vicinity of the proposed split seemed to fit equally well on either side. In such cases, rather than make an arbitrary decision, the career group in question was included in both clusters.

After clusters based on the hierarchical analysis had been tentatively determined, a final check was made on the basis of D^2 matrices (formula 7 data). Because the values appeared to be somewhat unstable for very small groups, the D^2 matrices were limited to career groups containing at least 50 cases, supplemented by a few small smaller groups that on the basis of the hierarchical analysis appeared to cluster with them. This resulted in limiting the number of groups in the D^2 matrix to 93.

Empirical results

The modified hierarchical procedure described in the previous section, in conjunction with the D^2 matrix procedure (formula 7) reduced the 93 career plan groups that were included (plus a 94th group: "undecided") down to 19 categories (plus a 20th for the "undecided" group). The categories are summarized in Table 2. Of the 19 categories, only 11 were clusters containing at least four groups. One contained just two groups and each of the remaining seven consisted of just a single career group that didn't cluster with anything else. (Among these seven unique and relatively homogeneous groups were architect and clergyman.)

As the opposite side of the same coin we have the handful of jobs that seemed to cluster naturally with more than one job family. A case in point is computer programer, which was the only career group falling in three separate job families. This unusual multiplicity of categories can probably be attributed to the fact that there are so many different kinds and levels of programers that one might almost say that the term "computer programer" doesn't denote any one job category.

Since we started with 173 career groups (in the hierarchical analysis) and ended with only 93 going into the 19 categories, what happened to the other 80? The answer is that each of these 80 groups had fewer than 50 cases, and some had fewer than 10. These groups, then, because of their small size, probably had rather unstable centroids. Consequently it still isn't entirely clear whether they are unique, clustering with no other group, or whether more data would fix their centroids so that it would become apparent that they belong in a cluster with other groups.

Table 3 shows the composition of each of the clusters, and Table 4 shows how homogeneous each cluster is, by presenting the within-cluster range of D^2 values, separately for the cognitive and noncognitive variables.

It is interesting to observe that clusters that were relatively homogenous in terms of the cognitive variables turned out to be fairly homogenous on the noncognitive variables too. As a further check, it is planned to investigate whether the Grade 11 data confirm the clusters established on the basis of the Grade 12 data.

But what significance are we to attribute to the fact that there were so many jobs that didn't fall in tight clusters and so few that did? Probably the basic significance of this outcome lies in all that is implied by the apparent nonexistence of a clear hierarchical structure underlying the career groups. The centroids of these career groups are to a great extent scattered widely in n-dimensional space rather than falling neatly in tight little clusters. The patterns of aptitudes and abilities that characterize various jobs are perhaps almost as diverse as the corresponding patterns for people. Therefore in selecting a career field, there should be less need for a square peg -- or even a scalene triangular peg -- to force himself into a round hole than there would otherwise be. The full range of jobs should include lots for scalene triangles of different shapes and sizes.

REFERENCES

[1] Flanagan, J.C., Dailey, J.T., Shaycoft, Marion F., Gorham, W.A., Orr, D.B., & Goldberg, I. Design for a Study of American Youth, 248 pp. Boston: Houghton Mifflin, 1962.

[2] Flanagan, J.C., Dailey, J.T., Shaycoft, Marion F., Orr, D.B., & Goldberg, I. Studies of the American High School. (Cooperative Research Project No. 226) Project TALENT Office, 1962. 375 pp.

[3] Flanagan, J.C., Davis, F.B., Dailey, J.T., Shaycoft, Marion F., Orr, D.B., Goldberg, I., & Neyman, C.A., Jr. The American High School Student. (Cooperative Research Project No. 635) Project TALENT Office, 1964. 738 pp.

[4] Flanagan, J.C., Cooley, W.W., Lohnes, P.R., Schoenfeldt, L.F., Holdeman, R.W., Combs, Janet, & Becker, Susan. Project TALENT One-Year Follow-Up Studies. (Cooperative Research Project No. 2333) Project TALENT Office, 1966. 348 pp.

[5] Johnson, S.C., "Hierarchical Clustering Schemes", Psychometrika, 32 (1967) 241-254.

[6] Shaycoft, Marion F., Dailey, J.T., Orr, D.B. Neyman, C.A., Jr., & Sherman, S.E. Studies of a Complete Age Group--Age 15. (Cooperative Research Project No. 566) Project TALENT Office, 1963. 370 pp.

[7] Shaycoft, Marion F. The High School Years: Growth in Cognitive Skills. (Cooperative Research Project No. 3051) Project TALENT Office, 1967. 376 pp.

[8] Ward, J.H., Jr. and Hook, M.E. "Application of an Hierarchical Grouping Procedure to a Problem of Grouping Profiles", Educational and Psychological Measurement, 23 (1963) 69-81.

[9] Ward, J.H., Jr., Buckhorn, Janice, and Hall, Kathleen, Introduction to PERSUB, PRL-TR-67-3(I), Lackland AFB, Texas: Personnel Research Laboratory, Aerospace Medical Division, August 1967.

[10] Ward, J.H., Jr., Hall, Kathleen, and Buckhorn, Janice, PERSUB Reference Manual, PRL-TR-67-3 (II), Lackland AFB, Texas: Personnel Research Laboratory Aerospace Medical Division, August 1967.

NOTES

[1] This research was conducted as a part of Project TALENT, a project conducted jointly by the American Institutes for Research and the University of Pittsburgh, pursuant to a grant from the U.S. Office of Education, Department of Health, Education, and Welfare. The opinions expressed herein, however, do not necessarily reflect the position or policy of the U.S. Office of Education, and no official endorsement by the U.S. Office of Education should be inferred.

[2] All formulas in this paper are in the Appendix, which also contains a section defining all the notation used.

[3] The author is indebted to Robert A. Bottenberg and Joe H. Ward, Jr., with whom she discussed the proposed analyses, for their helpful advice concerning hierarchical analysis; to Dr. Bottenberg for making available a copy of USAF Personnel Laboratory's GROUP 4 hierarchical analysis program [8,9,10]; and to Bradford W. Wade, who wrote a versatile hierarchical program (HIER) that used GROUP 4 as a starting point but was specifically designed for Project TALENT's needs, and who also wrote the series of auxiliary programs needed for preparing input to the hierarchical analyses and for computing the subsequent formula 7 matrices.

[4] All of the alternative formulas that were tried out are incorporated as options in the HIER program referred to in Note 2 above, as are many other formulas.

[5] Of the 4 alternative formulas, formula 14 gave results closest to formula 5 (and 7). Formula 13 gave the most dissimilar and least meaningful results. It did not work well with these data.

APPENDIX

I. NOTATION

n = no. of variables (= no. of dimensions)

g = no. of groups

N_j = no. of cases in group j

N = total no. of cases

$$N = \sum_{j=1}^{g} N_j \qquad (1)$$

y_{ijk} = score of individual k in group j on variable i

$$i = 1,2,3,\ldots n$$
$$j = 1,2,3,\ldots g$$
$$k = 1,2,3,\ldots N_j$$

$\bar{y}_{ij}$ = mean of variable i for group j

$$\bar{y}_{ij} = \frac{\sum_{k=1}^{N_j} y_{ijk}}{N_j} \quad (2)$$

s_{ij} = sample standard deviation of variable i for group j

$$s_{ij} = \sqrt{\frac{\sum_{k=1}^{N_j} (y_{ijk} - \bar{y}_{ij})^2}{N_j}} \quad (3)$$

σ_{ij} = estimate of population standard deviation of variable i for group j

$$\sigma_{ij} = s_{ij}\sqrt{\frac{N_j}{N_j - 1}} \quad (4)$$

P_{ijk} = special case of y_{ijk}, where the variables are principal components.

$d^2_{\overline{AB}}$ = squared distance between centroids of groups A and B.

$d^2_{\overline{(AB)}\,\overline{C}}$ = squared distance between centroid of combined groups A and B, and centroid of any other group C.

$D^2_{\overline{AB}}$ = square of distance between centroids of groups A and B, where each dimension is scaled so that the standard deviation of the distance between a point in group A and a point in group B is uniform for all dimensions. (This scaling has the effect of changing elliptical configurations to circular ones.)

σ^2_{A+B} = mean square distance of points in combined groups A+B from centroid of combined group.

$\sigma^2_{d_{AB}}$ = mean square distance between any point in group A and any other point in group B.

$\sigma^2_{w_{AB}}$ = within-group variance estimated on the basis of groups A and B.

σ^2_{A+B}, $\sigma^2_{d_{AB}}$, $D^2_{AB}\#1$, and $D^2_{AB}\#4$ are alternative bases for merging groups in the hierarchical analysis. (See formulas 11-14.)

II. FORMULAS

$$d^2_{\overline{AB}} = \sum_{i=1}^{n} (\bar{P}_{iA} - \bar{P}_{iB})^2 \quad (5)$$

$$d^2_{\overline{(AB)}\,\overline{C}} = \frac{1}{N_A + N_B}\left[N_A d^2_{\overline{AC}} + N_B d^2_{\overline{BC}} - \frac{d^2_{\overline{AB}}}{\frac{1}{N_A} + \frac{1}{N_B}}\right] \quad (6)$$

$$D^2_{\overline{AB}} = \sum_{i=1}^{n} \frac{\left(\bar{P}_{iA} - \bar{P}_{iB}\right)^2}{\sigma_{iA}^2 + \sigma_{iB}^2} \quad (7)$$

$$d_{\overline{AB}} = \sqrt{\sum_{i=1}^{n} (\bar{P}_{iA} - \bar{P}_{iB})^2} \quad (8)$$

$$D_{\overline{AB}} = \sqrt{\sum_{i=1}^{n} \frac{\left(\bar{P}_{iA} - \bar{P}_{iB}\right)^2}{\sigma_{iA}^2 + \sigma_{iB}^2}} \quad (9)$$

$$\sigma^2_{w_{AB}} = \frac{N_A \sum_{i=1}^{n} s^2_{iA} + N_B \sum_{i=1}^{n} s^2_{iB}}{N_A + N_B - 2} \quad (10)$$

$$\sigma^2_{d_{AB}} = d^2_{\overline{AB}} + \sum_{i=1}^{n} (s^2_{iA} + s^2_{iB}) \quad (11)$$

$$\sigma^2_{A+B} = \frac{1}{N_A + N_B - 1}\left[(N_A + N_B - 2)\,\sigma^2_{w_{AB}} + \frac{N_A N_B}{N_A + N_B}\, d^2_{\overline{AB}}\right] \quad (12)$$

$$D^2_{AB}\#1 = \frac{\sigma^2_{d_{AB}}}{2\sigma^2_{w_{AB}}} \quad (13)$$

$$D^2_{AB}\#4 = \frac{d^2_{\overline{AB}}}{\sigma^2_A + \sigma^2_B} \quad (14)$$

Figure 1. The Grouping Problem: What Groups Get Combined?

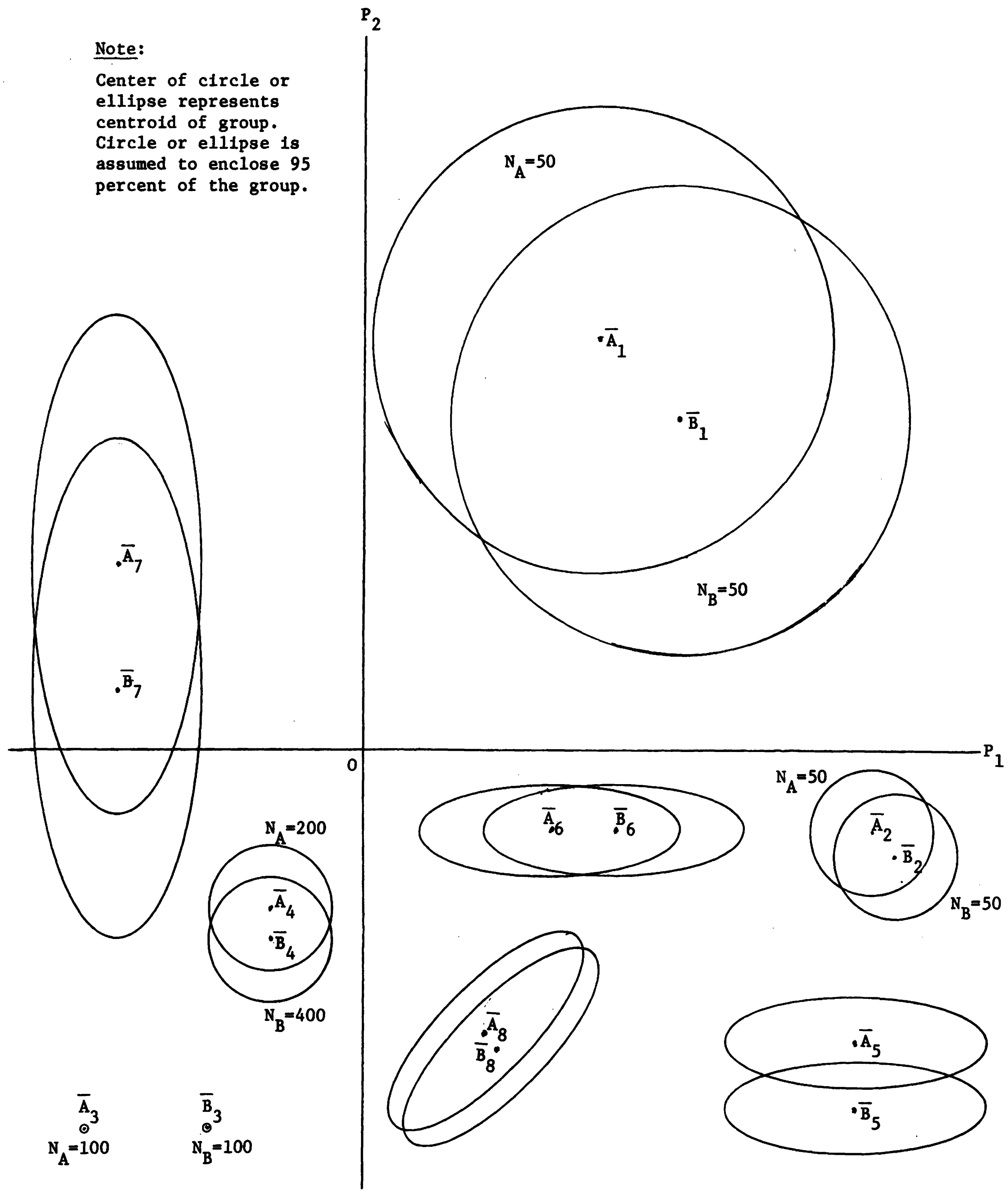
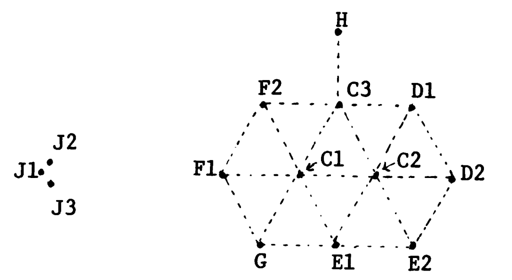


Figure 2. A Pitfall of the Ultrametric Maximum Method of Hierarchical Analysis: Empire-Building by Small Clusters at the Expense of Large Clusters

Figure 2a. A 14-point grouping problem



All line segments shown on this diagram are 2 units long. Except for the J and H points, all points are uniformly spaced, in a two-dimensional reticular pattern.

Figure 2b. Two intermediate stages in hierarchical analysis

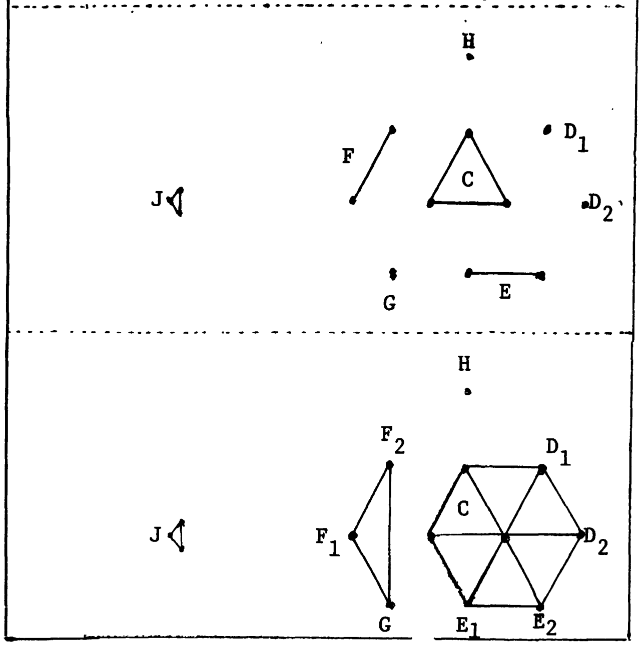


Figure 2c. Final grouping that would result from Ultrametric Maximum Method

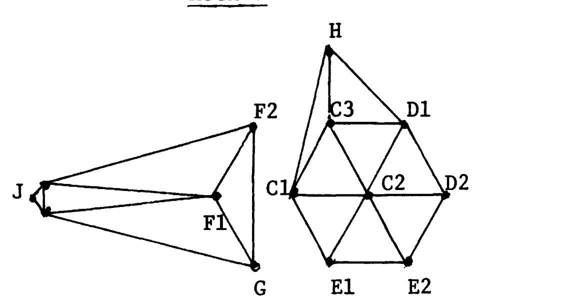


Figure 2d. Final grouping, using d^2 as the criterion

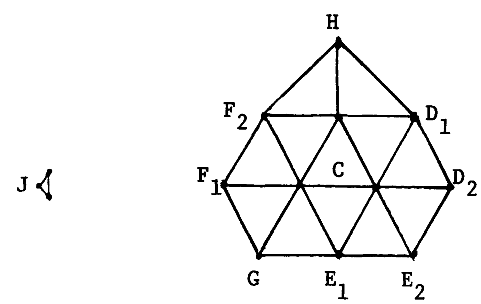
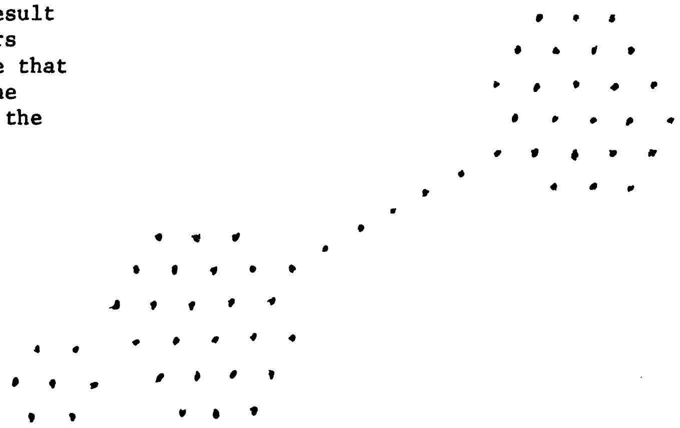


TABLE 1. Distance Matrix Corresponding to Figure 2

	C1	C2	C3	D1	D2	E1	E2	F1	F2	G	H	J1
C1		2.00	2.00	3.46	4.00	2.00	3.46	2.00	2.00	2.00	3.86	6.70
C2			2.00	2.00	2.00	2.00	2.00	4.00	3.46	3.46	3.86	8.70
C3				2.00	3.46	3.46	4.00	3.46	2.00	4.00	2.00	7.89
D1					2.00	4.00	3.46	5.29	4.00	5.29	2.83	9.85
D2						3.46	2.00	6.00	5.29	5.29	4.79	10.70
E1							2.00	3.46	4.00	2.00	5.46	7.89
E2								5.29	5.29	4.00	5.82	9.85
F1									2.00	2.00	4.79	4.70
F2										3.46	2.83	5.96
G											5.82	5.96
H												8.56

Figure 3. The Dumbbell Configuration: A Pitfall of the Ultrametric Minimum Method.

Applying the ultrametric minimum method of hierarchical analysis to the configuration shown at the right would result in the two clusters shown below. Note that the two ends of the "dumbbell" are in the same cluster.



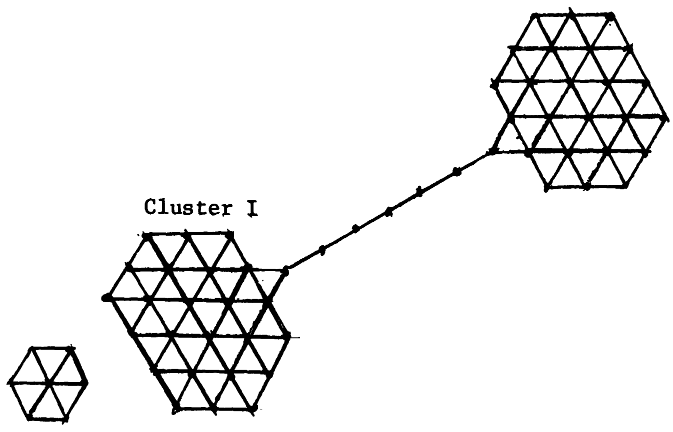
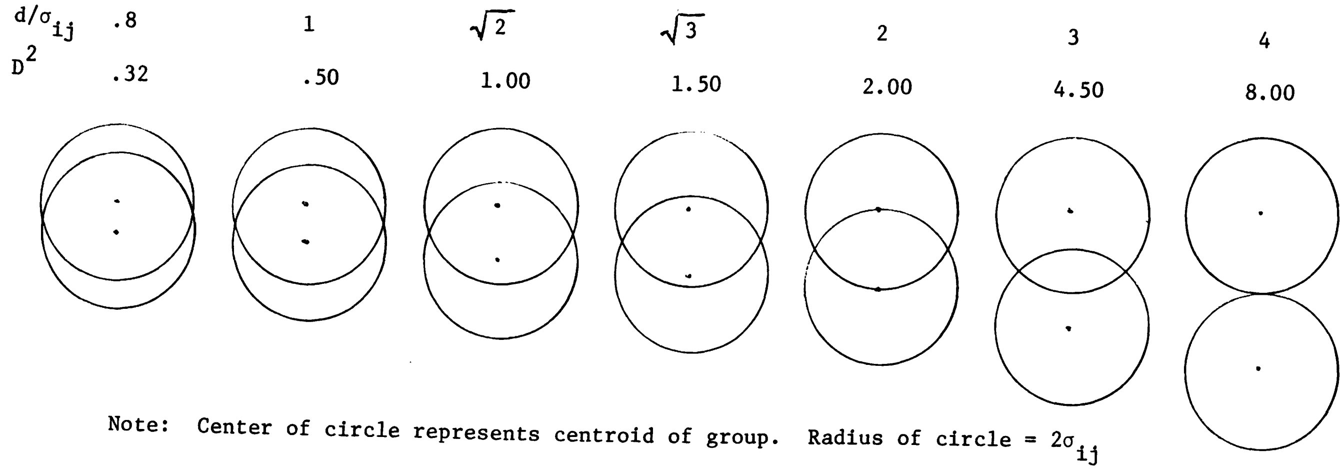


Figure 4: Amount of overlap corresponding to various values of D^2



Note: Center of circle represents centroid of group. Radius of circle = $2\sigma_{ij}$

TABLE 2. Job Families for Males*: Partial list (Based on 5-year follow-up of grade 12 boys)

	No. of career groups included	No. of cases included
A. Airplane pilot	1	99
B. Business and industry	19	2885
C. Architect	1	58
D. Engineering and applied physical sciences	12	1108
E. Math and physical science: Quantitatively oriented professions	5	287
F. Biological sciences: theoretical and applied	6	154
G. "People-oriented" professions in the sciences	4	500
H. Professions in the social sciences	4	763
I. College professor: English	1	57
J. Clergyman	1	176
K. Teaching and other "people-oriented" professions (non-science)	6	734
L. High school math teacher	1	62
M. High school science teacher	1	50
N. High school physical education teacher	1	57
O. Miscellaneous skilled occupations	7	673
P. Technician	6	610
Q. Miscellaneous "blue-collar" jobs	18	1543
R. Farming	2	372
S. Protective	4	259
T. (Undecided)	-	803
	100**	11250***

*Based on two hierarchical analyses of d^2 matrices (combining groups with the smallest d^2), with the resulting groups modified on the basis of the corresponding matrices of D^2 values. One of the d^2 matrices on which a hierarchical analysis was based was in 64 dimensions (principal components of 64 cognitive variables) and the other was in 45 dimensions (principal components of 45 noncognitive variables).

This table includes all groups having at least 50 cases (in the five-year follow-up of grade 12 boys). Some smaller groups are also included which fit into a family defined by the larger groups.

**Includes 7 duplications; therefore 93 separate categories.

***Includes 756 duplications; therefore 10494 separate cases.

TABLE 3. Present Composition of the Job Families

Job Family	Career Code	Career	No. of cases
A.	841	Airplane pilot	99
B.	112	In business for self (NEC)	278
	120	Industry or business (NEC)	141
	711(2)	Banking and finance	116
	716	CPA	280
	717	Accountant, auditor, comptroller (exc. CPA)	202
	723	Efficiency expert, industrial engineer, production management	111
	730	Business management, business administration (NEC)	511
	731	Manufacturing management	95
	732	Wholesale or retail trade management; marketing	195
	746	Insurance salesman	61
	748	Salesmen (NEC)	350
	749	Sales manager	73
	726	Personnel administration	81
	335	Pharmacist	91
	222*	Computer programer	121
	642*	U.S. Armed Forces: Officer	108
	747**	Auto salesman	22
	743**	Stockbroker	23
	745**	Real estate sales	26
C.	250	Architect	58
D.	240	Engineer (NEC)	167
	241	Civil and/or hydraulic engineer	105
	242	Electrical and/or electronic engineer	243
	243	Mechanical or automotive engineer	143
	231*	Chemist	81
	642*	U.S. Armed Forces: officers	108
	222*	Computer programer	121
	244**	Aeronautical engineer	37
	245**	Chemical engineer	58
	234**	Geologist	17
	235**	Meteorologist	10
	230**	Scientist (NEC)	18
E.	211	Mathematician	46
	232	Physicist	64
	231*	Chemist	81
	462*	College professor: Science	68
	461**	College professor: Math	28
F.	310	Biologist, zoologist, botanist, etc.	32
	314	Specialist in fish, wildlife, forestry, conservation, etc.	49
	313**	Specialist in agricultural science	35
	316**	Microbiologist	15
	317**	Biochemist	9
	322**	Surgeon	14
G.	329	Physician (NEC)	229
	332	Dentist	97
	360*	Psychologist	106
	462*	College professor: Science	68
H.	360*	Psychologist	106
	393	Lawyer	393
	463	College professor: social studies	85
	460	College professor (NEC)	179
I.	464	College professor: English	57
J.	521	Clergyman	176
K.	400	Teaching (NEC)	423
	420	High School Teacher (NEC)	69
	423	High School Teacher: Social Studies	94
	450	School administration (principal, etc.): except college	43
	370	Social work	60
	412	Teaching elementary school	45
L.	421	High school math teacher	62
M.	422	High school science teacher	50
N.	429	High school physical education teacher	57
O.	861	Printing trades	106
	758	Computer operator, etc.	78
	798	Miscellaneous administrative	38
	797	Miscellaneous clerical	112
	864	Draftsman	101
	738	Supervisor (in a business)	87
	661*	Police (public)	151

(Continued)

TABLE 3 (continued)

Job Family	Career Code	Career	No. of cases
P.	811	Electronic technician	212
	125	Electronics (NEC)	156
	299	Lab technicians, research assts., etc. (in physical sciences)	78
	222*	Computer programer	121
	347**	Medical and dental technicians; technicians in biol. and clinical sciences	16
	348**	Medical and dental technologists	27
Q.	102	Foreman (NEC)	49
	810	Electrician (NEC)	168
	812	Appliance repair	55
	820	Mechanic (NEC)	116
	821	Auto mechanic	74
	822	Airplane mechanic	68
	828	Machinist	176
	832	Carpenter	55
	833	Metal trades	92
	834	Bricklayer, mason roofer, printer, plasterer, etc.	62
	835	Plumber, pipefitter	72
	837	Misc. building and construction	71
	899	General labor (unspecialized)	277
	853	Auto, bus, and truck drivers	85
	824**	Industrial machine repair	21
	813**	Phone installation, repair, maintenance	39
	838**	Mining, quarrying, well-drilling	16
	836**	Operating earthmoving equipment; roadbuilding	47
R.	631	Farm or ranch owner	128
	639	Farming: other and miscellaneous	244
S.	661*	Police (public)	151
	666	Fireman	41
	640	U.S. Armed Forces (rank unspecified)	54
	641**	U.S. Armed Forces enlisted personnel	13
T.	001	(Undecided)	803

*Included in more than one group.

**Tentatively included in the group, on the basis of subjective decision, since the N is too small for conclusive empirical data.

TABLE 4. Range of Inter-career D^2 Values within Job Families

(Based on principal components*** of 64 <u>cognitive</u> variables or 45 <u>noncognitive</u> variables.)

Job Family	No. of Career Groups	No. of within-family D^2 values per matrix	Range of D^2 values (Formula 7)		
			Cognitive	Noncognitive	
A	1	0	----	----	
B*	16	120	.29-1.55	.18-1.82	Business and
B**	19	171	.29-3.80	.18-3.05	industry
C	1	0	----	----	
D*	7	21	.46-1.33	.40-1.29	Engineering and
D**	12	66	.46-6.15	.40-4.06	applied physical science
E*	4	6	.97-1.56	.83-1.31	Math and
E**	5	10	.97-2.59	.83-1.87	physical science
F*	2	1	1.57	1.27	Biological sciences
F**	6	15	1.57-6.09	1.27-6.49	
G	4	6	.79-1.30	.77-1.69	People-oriented, scientific
H	4	6	.54-.96	.63-.90	Social sciences
I	1	0	----	----	
J	1	0	----	----	
K	6	15	.47-1.81	.38-1.32	People-oriented, non-science
L	1	0	----	----	
M	1	0	----	----	
N	1	0	----	----	
O	7	21	.62-1.39	.59-1.59	Miscellaneous skilled
P*	4	6	.34-1.07	.25-.65	Technician
P**	6	15	.34-4.32	.25-2.76	
Q*	14	91	.41-1.60	.29-1.41	Miscellaneous
Q**	18	153	.41-4.03	.29-3.06	"blue-collar" jobs
R	2	1	.45	.20	Farming
S*	3	3	1.03-1.37	.84-1.64	Protective
S**	4	6	1.03-3.90	.84-3.94	
T	1	0	----	----	

*Excluding careers marked with double asterisk (**) in Table 3.

**Including careers marked with double asterisk (**) in Table 3.

***The principal components are based on scores of 14123 grade 12 boys who responded to the 5-year follow-up questionnaire. The information about careers was provided in the responses to that questionnaire.

ACE RESEARCH AND THE CONFIDENTIALITY OF DATA[1]

R. F. Boruch

American Council on Education

The Privacy Issue

Preserving the confidentiality of data has been the subject of much discussion and not a little controversy; numerous articles and books critical or defensive of current policies--have appeared. Westin's (1967) discussion of a wide variety of situations, including wiretaps, commercial records and psychological tests, is a scholarly treatment of privacy jeopardization. More recently, the proposal for a National Data Bank as described by Dunn (1967) and by Sawyer and Schecter, (1968) has led legislators and institutional researchers to give serious attention to the implications of such a project.

Within the social sciences, major concern about the issue has been given formal expression at a number of recent professional conferences. For example, at the 1968 American Personnel and Guidance Association meetings, a symposium, which included representatives of the American Council on Education, the Educational Testing Service, the National Opinion Research Center, and the National Merit Scholarship Corporation, addressed itself to confidentiality of data, specification of respondent rights, and the administrative problems in making data available to researchers. The American Educational Research Association has this year formed a standing committee, which will document and evaluate alternative approaches to existing problems arising from this issue. The American Psychological Association has initiated revision of its current professional code of ethics, based upon questionnaire survey of psychologists. The Russell Sage Foundation (New York) is devoting perhaps the most concentrated attention to the confidentiality question. Results of previous examination of issues under Russell Sage auspices are given by Reubhausen and Brim (1965); Goslin and Bordier are continuing work on educational administrative records.

Two attributes of most of these previous discussions are important in longitudinal research on higher education. First, many of them concern data collection installations which do not function as a source of data for evaluation. Records are used to create judgments about specific individuals. Such records must be highly reliable. Because longitudinal research data frequently has neither the same function, nor the same requirements for reliability, alternative approaches to maintain confidentiality can be developed; the import of this distinction will become evident when the system description is presented.

A second qualification to previous discussions is that most have been confined to justification of social science research in which respondent identification is necessary, and to endorsement of various ethics codes. Rarely do they specify methods for systematically assessing data collection with respect to assurance of confidentiality. Ethical codes are embraced but procedures for implementing these codes are rarely presented.

This paper is intended to clarify some of these issues and to suggest a useful approach to their examination. Using a rather primitive systems analysis, I have attempted to outline the ACE Cooperative Institutional Research Program (CIRP), and the way in which various stages and environments in the program can be assessed relative to the privacy issue. The function of the annual CIRP survey is longitudinal research on biographic, attitudinal and achievement attributes of college students.

ACE Policy and Procedures

Ethical guidelines are useful to the extent that they give to the interested participant or observer an unambiguous acknowledgement of the researcher's concern with safeguarding individual privacy: public misconceptions can be minimized. The delineation of ethics also serves as a useful reference system within which a policy may be implemented administratively.

In view of these considerations, the American Council on Education has formally encouraged the members of its research staff to support the code of ethics adopted by their respective professional organizations. The Council endorses adherence to the codes of ethics of, for example, the American Psychological Association and the National Education Association. When interpreted literally, each professional code includes some brief attention to the privileged nature of the respondent's relationship to the researcher and the obligation of protecting promised confidentiality of that relation. That such codes warrant improvement is apparent from recent professional conferences on the topic.

An explicit statement of policy relevant to administrative records developed and based on these codes of ethics was recommended by the American Council on Education (1967) to colleges and universities. In the statement, formation and implementation of clear policies for insuring confidentiality of students' records were recommended. Guidelines were presented in order to facilitate these recommendations (the use of legal consultation, the elimination of administrative records of political organization membership). Although the recommendations refer to administrative records specifically, they are applied by the ACE Office of Research to survey research records. In addition to these however, ther idealized regulations have been developed which evolve from the functional differences between administrative and survey records. That is, we (a) obtain records anonymously when research objectives do not include merges of data or follow-up studies (b) direct efforts toward safeguarding existing individual records such that identification by anyone (including ACE staff) are minimized or eliminated completely.

To help assure that these policies are implemented within the Office of Research, research activities are monitored by the Council's Board of

Directors, by the Office of Research Advisory Board, and by special advisory committees appointed for specific projects.

Administrative Policy and Procedures

In order to implement a reasonably good protection system, a wide variety of administrative devices have been incorporated into our operations. Regulations can be examined conveniently by using the flow chart given in Figure 1.

At the institutional level of data flow, the respondent group comprises all individuals within the participating institution who provide information about themselves during the questionnaire survey. Personnel at the institution are entirely responsible for administering and collecting questionnaires; the ACE Office of Research furnishes guidelines in order to expedite the process and detailed information is now being made available which describes the nature and functions of the Office of Research, research programs, etc.

The questionnaire contains a statement addressed to the respondent that briefly describes its research function and tells why identifying information is necessary. The respondent is also encouraged to cooperate in the research, under the acknowledgement of ACE responsibility for maintaining the confidentiality of the data. The American Council on Education has no authority to demand that the student respond to the questionnaire, although the institutional authority may indicate that he should complete it.

Under the direction of the institutional representative, the questionnaires are transmitted to a commercial service bureau for optical scanning and magnetic tape record creation. There they are maintained in locked files and destroyed when processing is completed. The product of the optical scanning operation is a statistical file which contains all responses and which uses arbitrary identification numbers for accounting purposes.

In a separate operation, identifying information is recorded on punched cards and then transferred to magnetic tape. This name-and-address file contains only identifying information and an accounting number. If name-and-address files and statistical information were matched, the total file would comprise an "intelligence system," as described by Dunn (1968), for example. However, additional coding at ACE comprises the basis for a double linkage protection system which prevents such matching even by the Office of Research personnel.

The double linkage system (currently being implemented) is illustrated schematically in Figure 2a. Each individual record in a given statistical data file is assigned a unique (arbitrary) accounting number. This series of numerals corresponds to Set 1 in the diagram. Each record in the corresponding name-and-address file is assigned another different accounting number (see Set 2). A code array (CA) of numbers, which match numbers in Set 1 to the corresponding one in Set 2 is created. The code linkage is maintained by a private organization under agreement to (a) allow no direct access to the code system to anyone, including ACE staff, and (b) merge existing accounting numbers with new ones. ACE copies of code linkage are destroyed. In order to implement follow-up studies (Figure 2b) more recent statistical data are assigned new accounting numbers (See Set 3). A new code linkage (CA') is then defined and translated by the service organization to the original system. Merges of statistical data occur without the problems involved in handling statistical and name-and-address files jointly. Accidental or deliberate disclosure of previously collected individual records is impossible simply because the linkage code is not in ACE possession and ACE personnel have no access to them.

At the ACE level of processing, the name-and-address files are maintained at a commercial service organization under the series of administrative constraints given in the Department of Defense Industrial Security Manual (1966) for "CONFIDENTIAL CLASSIFICATION." They are removed briefly from locked storage only for addressing follow-up questionnaires. Accounting controls include receipt and dispatch records, dates and time period of usage, and tape description. These administrative devices are maintained to insure against unauthorized tape copying.

The statistical files are not subject to the same rigid controls prescribed for name-and-address files. These data are so extensive that, even if one had full access to the statistical files and documentation, it would be virtually impossible to match individuals and their responses. Much the same is true for institutional identification.

The statistical data are consolidated and then summarized in various printed forms for the community of users outside the ACE Office of Research. Each institutional representative receives a statistical summary of responses to all questionnaires administered within his institution. This Institutional Report contains no information on individual respondents. To safeguard institutional privacy, ACE refuses to send to an institution another institution's report, although it does advise those researchers and administrators who want to compare their reports to contact each other directly. A second form of report, National Norms for Entering Freshmen (Creager et al., 1968), is a statistical summary of all data for a particular year and is provided to each participating institution and made available to the general public. It identifies neither individuals nor institutions. The only condition under which ACE statistical and identifying data are provided to the institution is met when the questionnaire has been administered under circumstances in which the student has been clearly informed in advance that the data would be returned to his institution for use in local research projects.

Current Problems and Alternative Solutions

At each level of the information system outlined there are potential difficulties in maintaining confidentiality of data. Consider the first level illustrated in Figure 1. Since identifying information appears on each questionnaire, this step represents a reduction in the ACE Office of Research control. That is, students or college personnel may have access to an individual record or a group of records. In cases where most questions are innocuous or useless for purposes other than research, the threat to the individual is of course, negligible. To the extent that college personnel and students endorse the principle of confidentiality and behave accordingly, the threat is not crucial. Thus far, most institutions acknowledge responsibility to treat document administration and collection confidentially.

More formal provisions for controlling questionnaire administration are possible. Physical security can be enhanced if respondents placed completed questionnaires in locked addressed boxes which would remain unopened until the data processing was initiated. Or the collection, and transmission of documents might be done under the surveillance of local student, faculty, and administrative representatives. More simply, the questionnaire could be constructed so that the identifying information is detachable from that portion of the document on which responses appear, and the identification section and completed questionnaires collected separately. Arbitrary identification numbers imprinted on both documents would permit later collation. The procedure, though not unwieldy, is expensive (approximately one-sixth higher than current processing costs). The types of controls mentioned may generally be too expensive, complicated, or time-consuming to be appropriate in most situations. Perhaps more importantly, very elaborate regulations and procedures could provoke mistrust or suspicion that would interfere with the research or with the operations of the institution. That is, if one implies that suspicion is warranted, then feelings of suspicion may increase or persist unnecessarily. So far, informal surveillance by local administrative personnel seems to be suitable for the general case.

The ACE policy has been to encourage and solicit cooperation by the respondent. The "voluntariness" of the situation is subject to modification by each college. There appears to be some justification for the college's requiring freshmen to complete the questionnaire, in that data are frequently collected in order to provide aid in planning future admissions policies, revising curricula, and other administrative decisions which affect the student. Insofar as the student feels that he cannot conscionably respond to questions relevant to social science or educational research, then conflict will arise. If the justification and restrictions on the research function are made clear to the respondent, conflict can be minimized and the student's cooperation may be forthcoming.

Questionnaire Processing Services

At the second level of the information system, that of the optical scanning of questionnaires and the production of magnetic tape records (name-and-address files and statistical data files), data may be misused, inasmuch as completed questionnaires usually include respondent identification. Irregular data usage at this level could take several forms, the most likely being reproduction of tapes for commercial exploitation.

If safeguarding magnetic tape record confidence is an important objective of the researcher, than endorsement of a relevant code of ethics by professionals in the computing disciplines is a reasonable expectation for reasons described earlier. Although members of the Association for Computing Machinery have discussed guidelines for professional conduct in data processing--some of them relevant to confidentiality--actual results are disappointing. While some members of this community recognize that an explicit code is desirable, other members appear to lack interest, and this indifference is one of the reasons that no code has been adopted. When translatable into procedures for safeguarding privacy of records, such a code can be included into contracts and so strengthen the efficacy of a desirable policy. There are some pitfalls in applying legal principles to technological environments (Banshaf, 1968), but legal applications appear to be a rather basic need if the matter of privacy is considered seriously.

Physical security in the data-processing environments could involve some sort of automated protection for tape or disc files. However, computer manufacturers acknowledge that little effort has been made toward developing devices which have data protection functions. There appear to be three major reasons for this lack of attention (Fanwick, 1967). First, competition among manufacturers is such that the development of such devices is not considered crucial unless there is a substantial demand for them. The demand is low, probably because of the social scientist's naivete regarding use of computing devices on the privacy context. Second, difficulties of undermining current administrative procedures and mechanical devices are sufficient to impede or discourage most attempts to misuse the data. Moreover, the costs required in development of new hardware-software devices may not result in systems which provide more protedtion than do current systems (Weismann, 1967). Third, it can be argued that in the data-processing environment, records are not available to persons without special skills and that this population of persons competent enough to misuse the data purposely is too small to justify anxiety.

The limited size and nature of the relevant community of potential violators is an important factor, since it means that adherence to existing, commonly used guidelines, can be monitored and controlled well. Further, by including liability and negligence clauses into contracts with individuals or organizations, effective incentives for continued attention to the privacy issue can be provided (Bigelow, 1969).

Office of Research Operations

At the ACE level of the information system, statistical information is analyzed and results disseminated to the public. In addition, follow-up studies are conducted, which require the use of name-and-address files. In addition to those mentioned above, various other safeguards for insuring confidentiality may be considered. The first depends largely on administrative regulation rather than on mechanical or automated procedures. An alternative device relates to computing and data processing methods. Yet another procedure involves capitalizing on the statistical nature of data analysis.

Consider, first, some of the problems that may be inherent in current administration procedures. It may be unwise to invest complete administrative control in a single individual. His personal attributes become too important and his absence, regardless of his character, may cause unnecessary inconvenience in operating the system. A "neutral" organization including members of respondent groups and of the interested professional community, might function as a surveillance or key control unit, in combination with the Director of Research. A second plausible alternative is to extend direct responsibility to other professional staff members of the Office of Research.

The first possibility entails considerable effort, as well as the cooperation of persons outside the ACE organization and so is not possible at present. Both alternatives have the disadvantage of being more bureaucratically complex and time-consuming than the current method.

The second device might involve an in-house computer system which permits merges of data while completely denying any individual direct access to internal name-and-address files and assigned accounting numbers. A model operation would include controls to eliminate direct handling of existing name-and-address files. Matching may be conducted independent of existing or newly obtained statistical files. The merge operations which combine new and existing statistical data are based on the accounting system associated with the existing name-and-address files. If such a system indeed prevents complete access during operations, its importance cannot be underestimated. By making direct linkage of name-and-address information and statistical data impossible, no attempt (legal or otherwise) could possibly threaten the confidentiality of the records and thus the privacy of a single person would no longer be an issue. However, a hardware-software system of this type is not currently available, although some systems can be developed to partially meet requirements. Relative to current administrative and mechanical devices, systems which completely eliminate accessibility would be considerably more expensive.

Consider now the third alternative, that which involves the form of the statistical data on which analyses are based. Typically, the researcher attempts to maintain an isomorphic relation between a person's responses on a questionnaire and records of these responses transformed to magnetic tape form. Now, the possibility of data use or misuse is, of course, weakened when data are not reliable for any <u>specific</u> individual record. Frequently, the researcher can afford to undermine deliberately the integrity of a single record but preserve the integrity of the whole, at least with respect to statistical parameters. He can do so by innoculating statistical data files with randomized error whose properties are known. A large body of literature deals with the problem of adjusting statistical estimates of population parameters, when the observations are subject to <u>known</u> measurement error. The innoculation accomplishes a number of important objectives. First, in the context of public interest in survey research, confusion between administrative records, eavesdropping devices, intelligence systems etc., may be minimized. The controlled unreliability of any individual record is a notion that can be communicated to the public. Second, the likelihood that records will be used in formation of judgments about specific individuals is reduced substantially. One cannot obtain unambiguous information about a specific person, even if identification is, infact, accomplished.

The procedure is inappropriate, of course, for detailed administrative records, but in the survey environment we are not so restricted. For many survey researchers within specific substantive areas (education, psychology, sociology) basic techniques (e.g., regression analysis, discriminant functions) for examining the structure of data can be augmented by including information available on measurement error parameters. Standard computer programs can be altered rather easily to adjust parameter estimates. Although nonparametric approaches which include misclassification probabilities not too well articulated, the current work on the topic ought to allow survey researchers some further latitude (e.g., Assakul and Procter, 1965).

Observations on the Legal Environment

Conceivably, public or private investigatory agencies may have an interest in an individual's responses to an ACE Office of Research questionnaire. Frequently, this information is already public. The researcher in possession of identical information is not confronted with a problem; the investigatory agency can obtain the data more efficently and easily from the public sources. The ACE Office of Research usually asks rather <u>general</u> information of the student, a strategy which not only protects against outside interference but also minimizes the intrusiveness of the questions. When solicited information is nonspecific, the risk of possible harassment to respondent and researcher is minimized but not eliminated.

Although the possibility that specific records survey data will be subpoenaed is a frequently mentioned bugaboo, the likelihood of this event actually happening is rarely assessed. Individuals who perceive such a threat frequently do so on emotional grounds rather than through systematic examination of the particular survey circumstance. It should be pointed out that (so far) in no instance has a subpoena of behavioral research survey data been effected. This lack of legal precedent can be interpreted in several ways. First, individual records directly relevant to investigatory objectives can usually be obtained easily through other agencies. Thus, an investigatory group has no need to solicit information which is of dubious relevance to its interests and which involves direct interference in social science research. The original survey questionnaires contain no signatures, and records of such documents are subject to well known psychometric limitations (i.e., unreliability of responses, and inaccuracy of data processing and of information transmission). In short, the survey research record is unlikely to have any value for use in litigation against a specific respondent.

The sociolegal aspects of the confidentiality issue are interesting and frequently paradoxical. Indeed, only recently has the research participant's right to privacy of personality been acknowledged by judicial and legislative action (see Reubhausen and Brim, 1965). For the researcher, the so-called privileged communication laws for psychologist-client relations appear to be relevant, but these regulations are statutory (extant in only 18states) and have been applied in rather limited situations.

Even though legislative or judicial inquiry into individual records is unlikely and even though strategies are employed to minimize private or public interest in such inquiry, the need for a legal definition of rights and obligations remains. Reubhausen and Brim, in a detailed examination of these questions, have offered suggestions for their resolution: One is that privileged status be extended to information acquired by the social scientist, another, that civil or criminal remedies for breach of the right of privacy be provided. Unfortunately, none of these suggestions are likely to receive immediate attention, evaluation, and action by the courts or legislative agencies. For the time being, heavy reliance must be placed on ethical codes, and on administrative procedures for their implementation within the particular survey condition. To these ends, the social scientist must devote serious effort to evaluating and implementing alternative strategies at each level of the information system.

References

Assakul, K. and Proctor, C.H. Testing Hypothesis With Categorical Data Subject to Misclassification. Institute of Statistics, Mimeograph Series No. 448, North Carolina State College, Raleigh, North Carolina: September, 1965.

Banzhaf, J.F. "When Your Computer Needs a Lawyer." Communications of the ACM. Vol. 2, No. 8, August 1968.

Bigelow, Robert P. "Some Legal Aspects of Commercial Remote Access Computer Services." Datamation. Vol. 15, No. 8, pp. 48-52. Chicago, Illinois: August, 1969.

Creager, J.A., Astin, A.W., Boruch, R.F., Bayer,A.E. "The National Norms for Entering College Freshmen-Fall 1968." ACE Research Reports. Vol. 3, No. 1, Washington, D.C.: American Council on Education, 1969.

Dunn, E.S. "The Idea of National Data Center and the Issue of Personal Privacy." The American Statistican. Vol. 21, No. 1, February, 1967, pp. 21-27.

Fanwick, Charles. "Computer Safeguards: How Safe Are They?" SDC Magazine. Vol. 10, 1967, pp. 26-28.

Ruebhausen, O.M., and Brim, O.G. "Privacy and Behavioral Research." Columbia Law Review. Vol. 65, 1965, pp. 1184-1211.

Sawyer, Jack, and Schechter, Howard. "Computers, Privacy, and the National Data Center: The Responsibility of Social Scientist." American Psycholgist. Vol. 23, No.11, November, 1968.

U.S. Department of Defense. Industrial Security Manual for Safeguarding Classified Information.

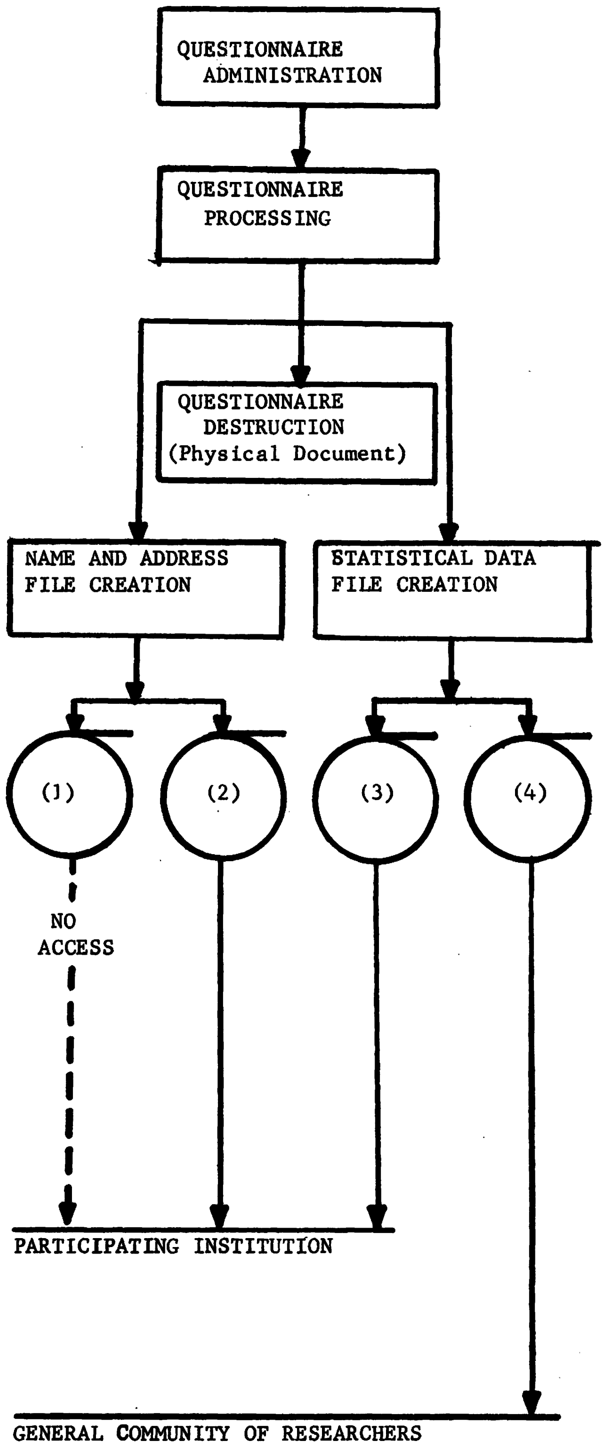
DOD 5220, 22-M. Washington, D.C.: Department of Defense, July 1, 1966.

Weismann, Clark. "Programming Protection: What Do You Want to Pay?" SDS Magazine. Vol. 10, No. 7, July, 1967, pp. 30-31.

Westin, A.F. Privacy and Freedom. New York: Atheneum, 1967.

Footnote

1. Partial support for this research was provided by NIMH Grant NM17084-01. Opinions expressed are not necessarily those of the sponsoring agency.



(1) DATA MERGE OPERATIONS
(2) STUDENT NAME AND ADDRESS FILE
(3) STATISTICAL (within college) REPORT
(4) NATIONAL NORMATIVE REPORTS

FIGURE 1. INFORMATION FLOW

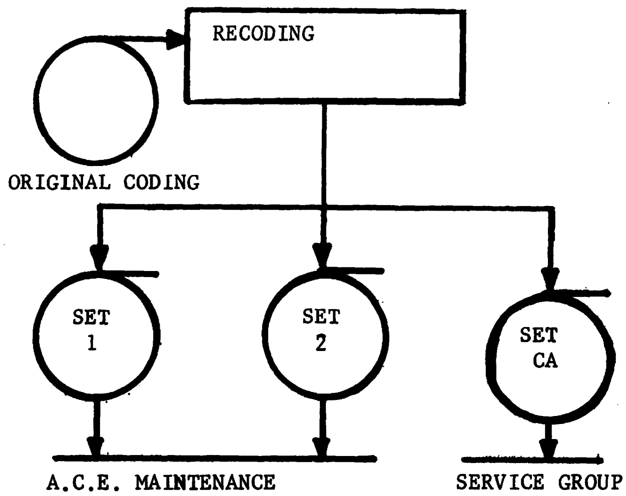


FIGURE 2a. LINK FILE CREATION

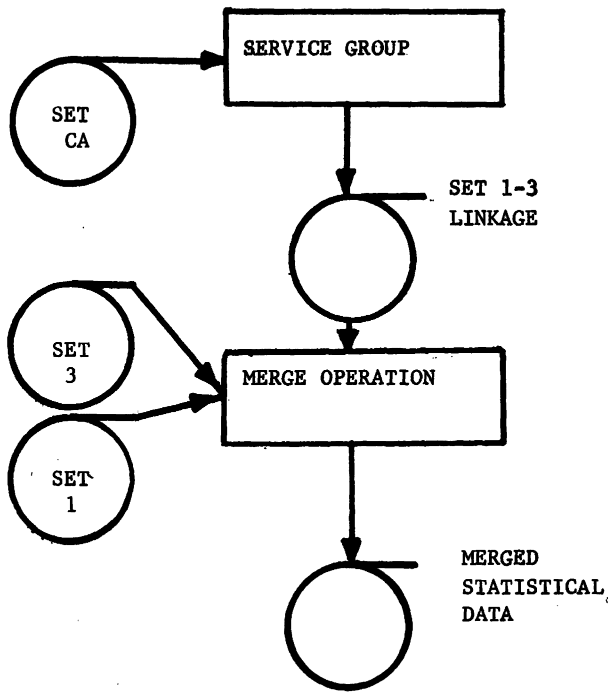


FIGURE 2b. MERGE PROCESS

THE IMPACT OF MILITARY SERVICE ON THE SUBSEQUENT CIVILIAN ATTAINMENT OF POST WORLD WAR II AMERICAN VETERANS*

William M. Mason, Graduate School of Business, University of Chicago

I have these findings to report about the effects of military service. The data are for men who served in the United States armed forces after World War II, but before the Viet Nam War.

First, and in general, military service makes a small contribution to veterans' social standing.

Second, this contribution appears to be greatest in providing or stimulating additional years of formal schooling.

Third, military service contributes to further attainment through military occupation, and through sheer length of service. That is, the higher the military occupation, the more additional school obtained. And, the longer a man serves, the more additional school he gets.

Fourth, the higher a man's military occupation, then the higher his civilian occupation as a veteran. However, length of military service is unrelated to civilian occupational status.

Fifth, neither military occupation nor length of service has anything to contribute to the income of veterans.

Sixth, I have found scant evidence to support the contention that the military has its greatest effects in raising the post-service social standing of veterans from poor backgrounds.

Data

The data used in this study are from the sample used in the May, 1964 Current Population Survey, and include all males then between the ages of 18 and 34. This means that the sample includes veterans who served after World War II and before Viet Nam. Furthermore, these veterans would have to be described as noncareer men, since the overwhelming majority of them served no more than four years.

For the purposes of analysis of the impact of service experiences, it is probably important that most of these veterans served in the peacetime military. If most had served during the Korean War, or during World War II, then an additional 10 per cent or so of them would have been in combat occupations, and this concentration in fighting specialties would have reduced the usefulness of the sample for studying crossovers between the civilian and military sectors. That is the chief advantage of a primarily peace-time sample--it maximizes the percentage of individuals in occupations which are not purely military. Furthermore, it is almost always with reference to a peace-time military force that policy discussions are carried on concerning the viability of the draft, opportunities for a second chance, and so on.

The conclusions I have already presented to you are based on a subsample of the veterans, who were 21 years of age or older at the time of the survey, who were also full-time employed members of the civilian labor force, and for whom relatively complete additional information was also available.

Method

To describe the effects of military service on veterans' social standing, multiple regressions were used with the subsample of veterans as the data set. These equations are identical in conception to those used by Blau and Duncan in their American Occupational Structure. That is, the equations taken together comprise a recursive system, in which the endogenous variables are identified as such by their choronological ordering. The present work may be thought of as adding an intervening step in the Blau and Duncan model of the process of social attainment. For present purposes, however, I am reporting the findings of only the last equations in the model, since the earlier equations do not bear on the effects of military service.

The variables for which the effects of military service were ascertained are: first, education during or after service. The data do not permit distinction between men who continued their education during service, and those men who continued their schooling only after service. The second variable for which the effects of military service were ascertained is occupational status--as measured by the Duncan occupational SES index. This variable is treated as endogenous with respect to education—regardless of whether the education was gained before, during, or after service. The third variable for which the effects of service were ascertained is occupational income. Finally, an alternative representation of income was also taken as endogenous--although, not with respect to occupational income. This variable is total income expected for the year, and differs from occupational income in that it includes other sources of income than the job held by the respondent at the time of the interview. Both of the income variables are taken to be endogenous with regard to educational attainment and occupational status.

Thus, the conclusions reported in this paper are based on four multiple regression equations. With respect to all that goes on in his life before a man is inducted into the armed forces, each of the equations contains the same regressors. The recursiveness comes simply from adding education during or after service into the equation for estimating occupational status after service, and then adding both of these variables into the equations for estimating occupational and total expected income.

The equations provide estimates of the effects of military service having also taken into account the effects of (1) father's occupation, (2) father's education, (3) the region the veteran grew up in, (4) the kind of place the veteran grew up in--for example, whether a large city, medium-size city, etc., (5) race--whether the respondent is white or Negro, (6) years of school completed before entering military service, (7) age, (8) score on the Armed Forces Qualification Test (AFQT), (9) whether the veteran had a full-time job before entering military service, and (10) the length of time since the veteran left military service.

This is a reasonably complete catalogue of variables the omission of which might bias the coefficients estimating the effects of military service.

The variables used to indicate the effects of service are (1) the length of time the man spent in service, and (2) the socioeconomic standing of his primary military occupation. The index of socioeconomic standing of military occupations is a scale which attempts to capture characteristics of military occupations. These characteristics are: the level of education found for men in a given military occupation, the paygrades of men in a given military occupation, the length of service school training for men in a given military occupation, the AFQT scores of men in a given military occupation, and the kinds of jobs held before service for men in a given military occupation. The reasoning used to construct the military occupational SES index is similar to the reasoning behind the Duncan occupational SES index for civilian occupations.

Note that my use of the military occupational SES index in the present context is to be distinguished from the analyses of other students of the military who create various scales of transferability between military and civilian occupations. The present index is not designed to measure equivalences between the civilian and military sectors, but simply to gauge the extent of knowledge, mental skills, rewards, and so on, which the military appears to require for each of its occupations.

The hypotheses associated with these variables are first, if military service helps to raise these noncareer military veterans' social standing, then it is reasonable to hypothesize that longer service leads to a greater increase in veterans' social standing. Furthermore, if men gain skills in their primary military occupations which are useful in civilian life, then insofar as these skills are associated with the standing of their military occupations, it follows that the higher their military occupations, the higher their civilian standing as veterans.

Findings

Table 1 and Figures 1 and 2 present the evidence concerning the effects of length of service and military occupational SES on subsequent veterans' attainment.

Table 1 presents a decomposition of the multiple correlation for each of the four equations estimated. This decomposition is, of course, one of many possible. The results, however, are so clear that it hardly matters which decomposition of the explained variance one uses.

Table 1 shows that length of service and military occupational SES are but minor contributors in estimating additional years of school gained during or after service. Together, they account for only 16 per cent of the explained variance. By contrast, education before service, and the AFQT test--which may be treated as an estimate of quality of education, account for nearly half the explained variance.

The second equation estimates veterans' occupational status. This time, military service accounts for only 4 per cent of the variance. By contrast, education, and AFQT score account for over half of the explained variance. Note, however, that schooling obtained during or after service is also a sizable contributor to occupational SES. Hence, military service makes an indirect, as well as a direct, contribution to veterans' occupational standing.

Finally, as Table 1 indicates, the military service variables used here make no direct contribution to income.

Figures 1 and 2 present this information in a different way. They are path diagrams in which the effects of variables preceding military service for simplicity have not been explicitly included. But, the path coefficients--that is, the standardized multiple regression coefficients, are in fact computed net of the background variables. Figure 2 differs from Figure 1 only in choice of the income measure as the last endogenous variable. A comparison of the two figures indicates that the indirect effects of military service are smaller for expected yearly income than they are for occupational income. This results from the weaker association between education gained during or after service and expected income.

In sum, as measured here, military service affects years of school attained after entering service, and it affects veterans' occupational SES. Its effects on income are entirely indirect. Those who would argue for the military's salutory effects on the occupations and income of its veterans will not find overwhelming support of their hypothesis in this data. It may be that military experiences do indeed lead to higher social positions for veterans, but to the extent that this is so, it occurs because of increased education, more than anything else.

The argument may be modified to say that the military contributes more to the social standing of those who start out their lives in a low position. This has been one of the positions taken by various government officials in the recent past. The 1964 veterans data were examined to find support or denial of this hypothesis. In the regression framework, the argument can be tested through the use of an interaction model. This interaction model is a simple operationalization of the hypothesis that men of low social origins get more of a boost from the military than do others.

The computational process is to define a subset of veterans who were of low social origins, and then proceed to determine whether this additional knowledge is useful. More precisely, I defined a dummy variable which takes the value one if the veteran was below the mean on all salient background variables, as well as below the mean in achievement before service, and on the Armed Forces Qualification Test. The variable takes the value zero otherwise. This dummy variable was then allowed to interact with the military variables.

A final variable taken into consideration for this part of the analysis is the product of military occupational status with a dummy variable which distinguishes between men who had a job before service, and those who did not. There are competing hypotheses about the directionality of the coefficient associated with this variable; but for present purposes it matters only that inclusion of this variable allows for a more rigorous specification of the model.

Table 2 presents squared multiple correlations for the equations without the interaction terms, for the corresponding equations with the interactions, and presents the F-ratios associated with the increments in the sums of squares accounted for. The F-ratios under these circumstances have the status of a bench mark, a handy reference point, since the sample and data only begin to approximate the conditions needed for use of the F-distribution. The results are clear: including the interaction terms in the equations has essentially no effect on the sums of squares explained. As shown earlier, military service has its greatest effect on gaining additional years of school. The interactions add nothing to the explanation of this variable. And, although the F-ratio for the increment in the explained variance for occupational SES gets its nose above water, it does no more than that. In sum, the data and method used here do not support the hypothesis that, to reverse a well-known phrase, the "poor get more" out of service than those of higher social origins.

I would like to stress that the data I have been working with do not include many of the men falling in the lowest mental groups, as defined on the distribution of the AFQT scores. The armed forces, and primarily the Army, are currently carrying out a program to induct and train men who in the past would have been rejected by the services because of their low AFQT scores. This program, called PROJECT 100,000 may yet prove more successful than any extrapolation from the finding reported here might indicate. Furthermore, it must be pointed out that men likely to fall under the aegis of PROJECT 100,000 are only marginally in the labor force. A major contribution for the services for these men might be simply to enable them to find and keep jobs, low status though these jobs might be. I have been unable to test this potential effect of military service with the present data because of insufficient observations.

Other Points

I would like to turn now to a brief consideration of additional information, pertaining to the armed forces and education. One interesting question is, when does a serviceman go back to school, if he does at all? And, given that he goes back to school, at what point in the education distribution is he likely to be located? Without attempting anything like a comprehensive answer to these questions, Table 3 presents the distributions for education before service, and current total education, for enlisted men who were in the armed forces in 1964, at the same time that the veteran and nonveteran data were collected. The table shows that 21 per cent of the men in one distribution would have to be redistributed in order to make the two distributions identical. Very clearly, much of the effect of service on education takes place within service, and not afterwards.

A second point to be gleaned from this table is that the principal impact of the services on educational attainment is in helping men to complete high school.

Finally, I have carried out a simple comparison between veterans and nonveterans. This comparison is done by means of indirect standardization, which requires no scaling or other assumptions about the variables or relationships. What I have done is to compare veterans and nonveterans by their occupational distributions. For the veterans, I have defined four groups, as determined by the cross-classification of whether a man completed his education before service or not, and whether the man entered the labor force before service or not.

Table 4 presents the occupation distributions for each of these groups of veterans, as well as all groups aggregated, and for nonveterans. The table indicates considerable differences in occupation distributions for the groups: veterans who went back to school are more likely to be white collar than nonveterans, or veterans who did not go back to school. In particular, about 36 per cent of nonveterans are white collar workers, as compared with about

35 per cent white collar for veterans who did not go back to school during or after service, and as compared with about 58 per cent white collar for veterans who did go back to school during or after service.

Using the combined veterans and nonveterans sample as the population base, the occupation distributions were indirectly standardized for age and education. The results, which are presented in Table 5, indicate that age and education account for much of the differences observed in the unstandardized distributions. As a result of the age and education standardization, per cent white collar for any of the veterans groups and nonveterans range within 5 per cent of each other. Taking entire distributions into account, rather than simply per cent white collar, provides the same finding. Table 6 presents indices of dissimilarity between the various distributions. Unstandardized, these indices are rather large; standardized, they drop considerably, particularly when comparing nonveterans with the veterans groups.

The standardizations indicate, then, that much of the differences in occupational attainment between veterans and nonveterans can be explained by age and education. And this is wholly consistent with the findings I started out with--military service appears to make its greatest contribution via additional years of school.

I should like to indicate my awareness that this simple comparison of veterans and nonveterans hardly exhausts the subject. I expect in the near future to carry out more detailed comparisons of veterans and nonveterans.

These results are of interest because they help to delineate the way in which a major institution of this country intervenes in the life cycle. The results also bear on policy discussions which require information about the consequences of service. They may also be taken as the preliminary skirmishes for a human capital analysis of the net value of military service for the men who serve.

---

*This research was supported in part by a grant from the Inter-University Seminar on Armed Forces and Society.

TABLE 1

Percentage of Explained Variance Due to Some Aspects of Military Service, and Certain Other Variables, in Regression Equations Estimating Selected Characteristics of Veterans' Social Positions[a]

Independent Variable	Dependent Variable			
	Additional Years of School During or After Service	Log Current Occupational SES	Log Current Occupational Income	Log Expected Total Yearly Income
Years of school completed before entering service, and Armed Forces Qualification Test score..........	44%	53%	21%	13%
Socioeconomic status of military occupational specialty............	4	4	0	1
Length of active duty military service..........................	12	0	0	0
Additional years of school during or after service.....................	...	17	7	1
Log of socioeconomic status of current occupation................	...	...	23	25
All other variables in equation[b]..	40	26	49	60
Total[c]..........................	100%	100%	100%	100%

[a]Data source is a veterans' subsample (N = 1450) of a 1964 Current Population Survey of civilian men between the ages of 18 and 34.

[b]Includes the socioeconomic status of father's occupation, father's education, region the veteran grew up in, kind of place the veteran grew up in, race, whether the veteran had a full-time job before entering military service, the length of time since the veteran left military service.

[c]Decomposition of explained variance computed according to the identity:

$$100 = \left(\frac{\sum_{i=1}^{k} b_i^* r_{yi}}{R^2} \right)(100)$$

TABLE 2

Squared Multiple Correlations, for Regression Equations Estimating Veterans Social Standing, with and without Variables Distinguishing Veterans With Low Background Standing from Other Veterans[a]

Dependent Variable	R^2 without Interactions	R^2 with Interactions[b]	F for increment in R^2
Additional years of school after service.................	0.165	0.168	1.3
Log of socioeconomic status of current occupation............	.295	.302	3.6[c]
Log of current occupational income........................	.244	.245	0.4
Log of expected total yearly income........................	.216	.218	0.9

[a]For description of variables in equations without interaction terms, see Table 1.

[b]Additional variables included in these equations are as follows:

z_1 = 1, if veteran's father went no further than 8th grade, if veteran's father's occupational SES was no higher than 25, if veteran had no more than 3 years of high school before service, and if veteran's AFQT score was no higher than 50; z_1 = 0 otherwise.

z_2 = product of z_1 with log military occupational SES.

z_3 = product of z_1 with duration of active military service.

z_4 = product of log military occupational SES with a binary variable which takes the value 1 if veteran had a full-time job before entering service, and takes the value 0 otherwise.

[c]$F_{0.95} = 2.4$ for (4,1000) degrees of freedom.

TABLE 3

Percentage Distributions of Years of School Attained Before Entering Military Service, and Total Years of School Attained, for a Sample of the Enlisted Ranks of the Armed Forces in 1964[a]

Education Level	Years of School Completed[b]	
	Before Service	Total
Less than eighth grade	2.3	0.8
Eighth grade	6.0	2.9
Ninth to eleventh grade.	33.2	16.4
High school graduate	44.3	57.8
One or two years of college.	9.3	15.5
Two or more years of college, but less than a college B.A.	3.9	5.2
College B.A.	0.9	1.1
Study beyond the college B.A.	0.2	0.3
Total of percentage distribution	100.1	100.0

[a]Data for this table are from a 1964 Department of Defense survey of all branches of the military for all ranks up to O-6. This table is based on a subsample (N = 48,123) of the total sample.

[b]Index of dissimilarity between the two distributions is 21. The index of dissimilarity is the minimum percentage of either distribution which would have to be redistributed so that both distributions were identical.

TABLE 4

Percentage Distribution Over Current Full-Time Occupation, For Non-Veterans, All Veterans, and Four Sub-Groups of Veterans; All Men Aged 18-34, and Sampled in Autumn 1964[a]

Occupation Group	Group[b]					
	Non-Veterans	Veterans	Veterans Group I	Veterans-Group II	Veterans-Group III	Veterans-Group IV
Professional, technical, and kindred workers.....	13.7%	14.5%	9.8%	23.0%	8.3%	31.7%
Farmers and farm managers............	2.0	1.2	1.9	. . .[c]	1.4	. . .[c]
Managers, officials and proprietors, except farm........	6.2	11.0	9.8	13.2	10.4	13.3
Clerical and kindred workers....	7.6	8.4	7.8	9.0	8.8	8.8
Sales workers.......	6.3	6.6	4.9	8.6	7.4	8.3
Craftsmen, foremen, and kindred workers.............	16.2	22.4	26.5	20.0	22.1	13.3
Operatives and kindred workers.....	26.1	24.6	28.3	17.4	27.1	16.3
Service workers, except private household...........	7.5	5.8	5.2	4.6	7.9	4.5
Farm laborers and foremen.............	4.0	1.1	1.4	0.2	1.5	. . .[c]
Laborers, except farm and mine.......	10.3	4.5	4.4	3.9	5.1	3.7
Total percentages and case bases......	99.9% (3970)	100.1% (2651)	100.0% (1127)	99.9% (409)	100.0% (740)	99.9% (375)

[a]Source for this table is the CPS survey conducted in November 1964 for the Department of Defense Study of the Draft.

[b]Defined for those men who reported that they were employed full time. The "All Veterans" is decomposed into four sub-groups defined as follows: Group I consists of men whose first full-time job came before service, and who completed their education before entering service. Group II consists of men whose first full-time job came before service, and who completed their education during or after military service. Group III consists of men whose first full-time job came after service, and who completed their education before entering service. Group IV consists of men whose first full-time job came after service, and who completed their education during or after military service.

[c]No observations fell into cell.

TABLE 5

Percentage Distribution over Current Full-Time Occupation, for Non-Veterans, All Veterans, and Four Sub-Groups of Veterans. All Men Aged 18-34 and Sampled in Autumn 1964--Standardized for Age and Education, Using the Entire Veterans and Non-Veterans Sample as the Base[a]

Occupation Group	Group					
	Non-Veterans	All Veterans	Veterans-Group I	Veterans-Group II	Veterans-Group III	Veterans-Group IV
Professional, technical, and kindred workers.....	14.9%	12.7%	14.8%	12.7%	8.5%	14.8%
Farmers and farm managers...........	2.2	1.0	1.3	0.2	1.2	0.4
Managers, officials and proprietors, except farm........	7.9	8.4	8.5	7.6	8.8	8.3
Clerical and kindred workers....	7.7	8.2	8.4	7.3	8.7	7.4
Sales workers......	6.3	6.7	6.4	6.5	7.7	6.3
Craftsmen, foremen, and kindred workers...........	17.7	20.3	21.6	21.1	19.5	17.3
Operatives and kindred workers....	24.6	27.0	26.2	27.0	27.6	27.9
Service workers, except private household.........	6.9	6.8	5.7	6.5	8.6	6.8
Farm laborers and foremen...........	3.3	2.1	1.7	2.3	2.4	2.0
Laborers, except farm and mine.....	8.7	7.0	5.5	8.7	7.0	8.8
Total percentages[b]	100.2	100.2	100.1	99.9	100.0	100.0

[a]The technique of standardization used is that of "indirect standardization." The standardizing variables are age (grouped: 18-21, 22-25, 26-29, 30-34) and total education (grouped: eight grade or less, some high school, high school diploma, some college, and college B.A. or more schooling). The standard population used in the analysis is the combined veteran and non-veteran sample appearing in Table I. For a discussion of standardization techniques, see Evelyn M. Kitagawa, "Standardized Comparisons in Population Research," Demography, Vol. 1, No. 1, (1964), pp. 296-315, esp. 312 ff.

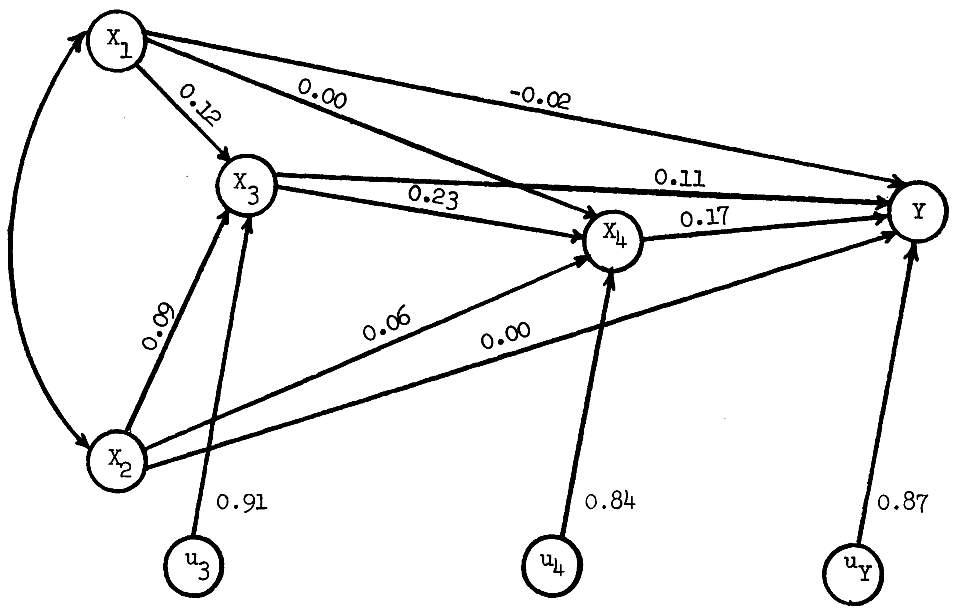
[b]Case bases for this table are necessarily identical to those for Table I.

TABLE 6

Indicies of Dissimilarity between Distributions Listed in Tables 4 and 5[a]

Variable	Non-Veterans	All Veterans	Veterans Group I	Veterans Group II	Veterans Group III	Veterans Group IV
Nonveterans	4.7	12.9	16.2	23.8	13.7	28.3
All veterans	6.4	6.9	8.9	13.4	7.0	21.7
Group I veterans	6.9	4.0	8.9	21.5	7.6	29.8
Group II veterans	6.2	2.9	5.6	19.7	18.9	8.8
Group III veterans	9.9	4.9	8.4	7.6	4.5	27.3
Group IV veterans	3.9	4.9	6.5	4.3	8.4	25.3

[a]Entries above the diagonal are for unstandardized distributions. Entries below the diagonal are for standardized distributions. Entries on the main diagonal are for the same variable in its standardized and unstandardized forms.



Identification of Variables

X_1 = length of active duty military service

X_2 = socioeconomic status of primary military occupation

X_3 = additional years of school during or after service

X_4 = log of socioeconomic status of current occupation

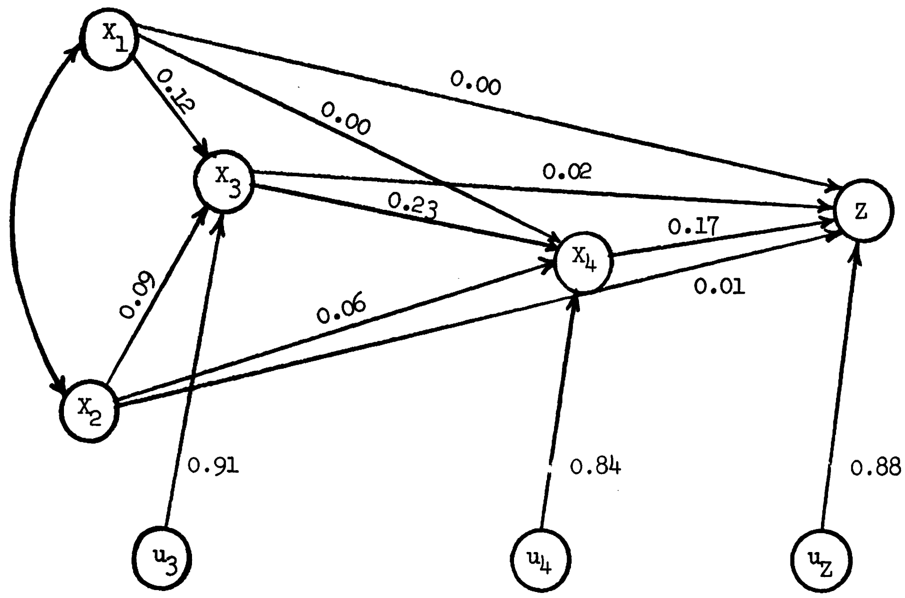
Y = log of occupational income

u_3 = disturbance term in estimating X_3

u_4 = disturbance term in estimating X_4

u_Y = disturbance term in estimating Y

Fig. 1.--Path diagram illustrating effects of military service on veterans' subsequent social standing using occupational income; extracted from equations indicated in Table 1.



Identification of Variables

X_1 = length of active duty military service

X_2 = socioeconomic status of primary military occupation

X_3 = additional years of school during or after service

X_4 = log of socioeconomic status of current occupation

Z = log of total expected yearly income

u_3 = disturbance term in estimating X_3

u_4 = disturbance term in estimating X_4

u_Z = disturbance term in estimating Z

Fig. 2.--Path diagram illustrating effects of military service on veterans' subsequent social standing using total expected yearly income; extracted from equations indicated in Table 1.

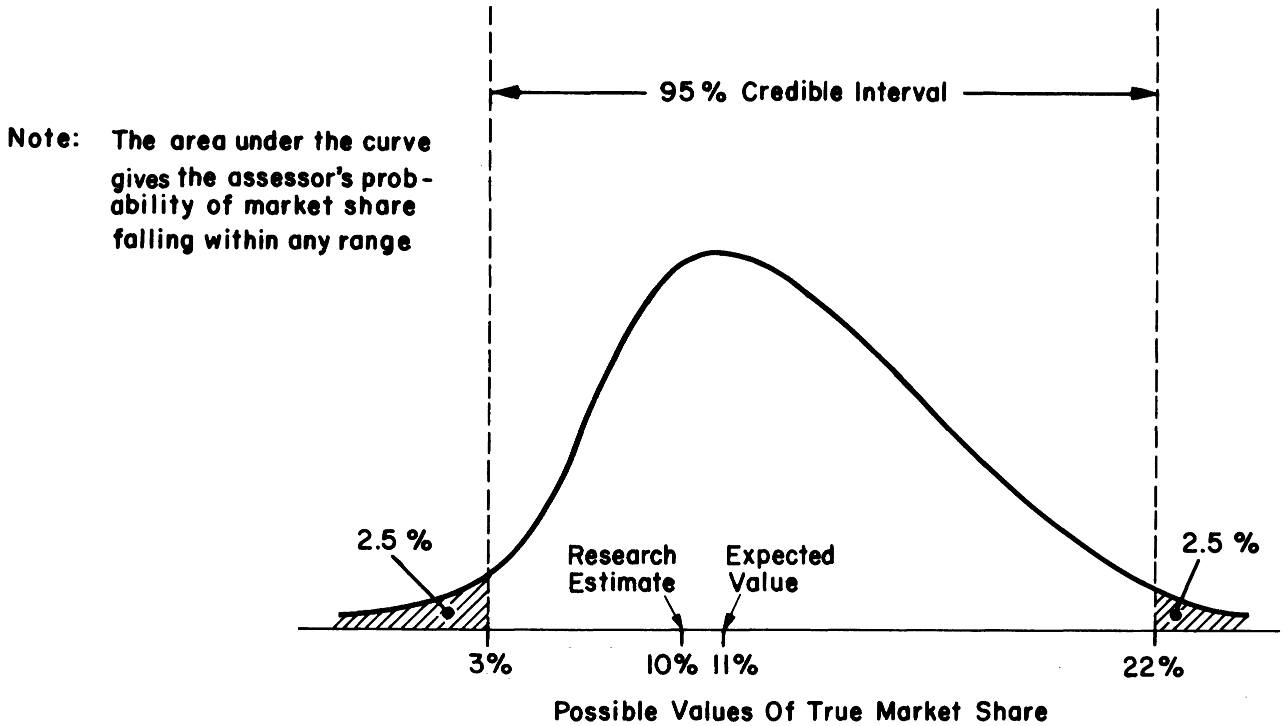


FIG. 1 REPRESENTING PERSONAL UNCERTAINTY

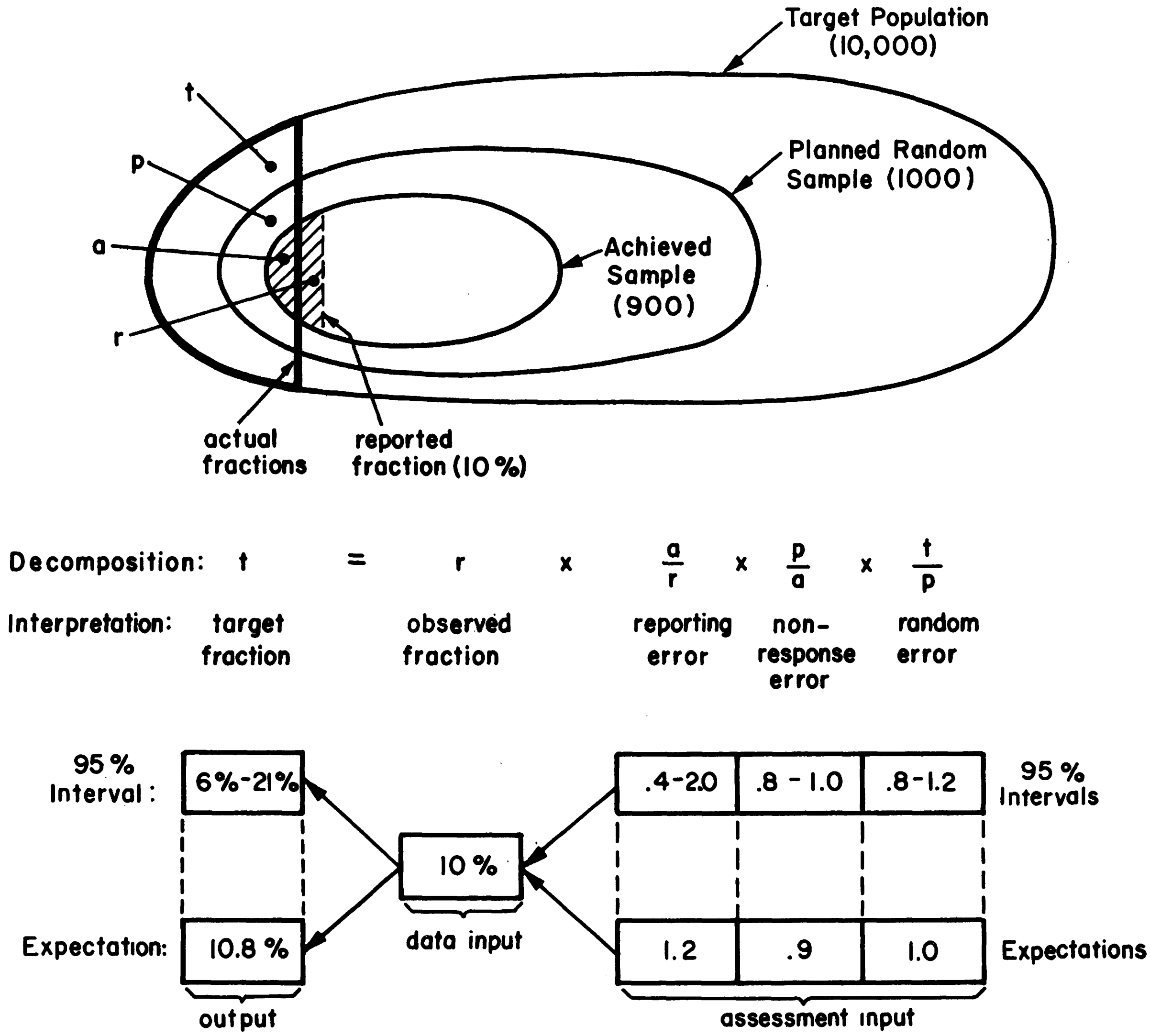


Figure 2. ASSESSING A POPULATION FRACTION FROM A SURVEY

XVI

CONTRIBUTED PAPERS VI

STATISTICAL SURVEY METHODS: NEW APPLICATIONS AND DEVELOPMENTS

Chairman, CARL L. ERHARDT, New York City Health Services Administration

THE BIAS OF THE ORDINARY LEAST SQUARES ESTIMATOR OF THE REGRESSION COEFFICIENT FOR A BIVARIATE POPULATION WHEN BOTH VARIABLES ARE SUBJECT TO CORRELATED MEASUREMENT ERRORS*

John J. Chai, Syracuse University

1. INTRODUCTION

In surveys of socio-economic status or in scientific experiments, the influence of measurement errors always exists.

A stock boy taking an inventory of certain products at a specified time and day by counting the number of items for every product is expected to make counting errors. These counting errors may or may not be correlated between different units he counts on that day. The counting errors may be correlated between the units which are counted in a given day, if the stock boy counts the items using a certain method in the morning, say, and then he finishes counting the items using another method in the afternoon. Reporting on the number of family members by self enumeration of a respondent or by an interviewer produces another type of measurement errors, response errors. It has been well known from past studies that response errors of this kind are correlated within an interviewer assignment area due to interviewer bias in consecutive interviews. Even when a self-enumeration method (where no interviewer bias is involved) is used, we expect correlated measurement errors. For example, consider a survey of price of houses in a community. Suppose that people in the community are asked to assess their own homes. If a person in that community had just sold his house at a certain price level, the assessment of the other houses in the community may be affected [10].

Correlated measurement errors may also be expected in measuring the length or weight of an object, or in consecutive readings of fluctuating temperature of an instrument in a chemistry laboratory, or in grading student papers. For example, consider the grading of papers by an instructor. If the instructor grades one group of papers at one time, rests, and then grades the remainder of the papers at another time, we expect errors in grading to be correlated.

Theory of measurement errors in sample or census surveys for univariate case has been developed for some basic survey conditions in the recent past [1, 2, 4, 8, 9, 10, 13].

So far, we have illustrated the cases in which a single variable is taken separately. Now consider an example for a bivariate case where the characteristics of interest are the height and weight of a person. Measurement errors in this case may be caused by either instruments or by the person (or persons) measuring height and weight or by both. The measurement errors associated with height may be correlated with the errors associated with weight, positively or negatively. Furthermore, measurement errors for each variable may be correlated between the units within a set of observations. Therefore, if a person is interested in studying the relationship between two variables, he should recognize the existence of measurement errors and their effects on the estimators of relevant parameters. For example, suppose that we measure a set of bivariate characteristics X and Y and that the measurement is made with errors. Let the observed values (what we actually measure) be denoted by x and y and the measurement errors associated with X and Y by d and e respectively. Suppose further that

$$x = X + d \tag{1.1}$$

$$y = Y + e \tag{1.2}$$

and that the relationship between the two variables when neither variable is subject to measurement errors can be described by a linear functional model[1/]

$$Y = \alpha + \beta X \tag{1.3}$$

However, the observed variables are x and y, so the model (1.3) above becomes

$$y = \alpha + \beta(x-d) + e \tag{1.4}$$

by (1.1) and (1.2). And if we let

$$\omega = e - \beta d \tag{1.5}$$

Then (1.4) becomes

$$y = \alpha + \beta x + \omega \tag{1.6}$$

One may be interested in estimating β to see whether there exists a statistical relationship between X and Y from a set of sample observations. Suppose that the person tried the ordinary least squares estimator from a sample of n observations

$$b = \frac{\sum^{n}(x-\bar{x})(y-\bar{y})}{\sum^{n}(x-\bar{x})^2} \tag{1.7}$$

where $\bar{x}$ and $\bar{y}$ are sample means of x and y, respectively. It is well known [12] that the estimator given by (1.7) is biased and inconsistent, i.e., Plim b = $\beta/(1+\sigma_d^2/\sigma_X^2)$. Thus b underestimates β unless $\sigma_d^2 = 0$, even when the errors d e are mutually and serially independent with constant variances, σ_d^2 and σ_e^2 and are also independent of X and Y.

There has been a considerable amount of theoretical work done in the past in developing better estimators (e.g., consistent estimators) of β when both variables are subject to errors (e.g., [3], [6],[7],[11],[12],[13],[15]). However, the ordinary least squares estimator is used more than often in practice, whether the variables concerned are subject to measurement errors or not. And there are many survey or experimental situations, where the ordinary assumptions mentioned earlier (the mutual and serial independence of errors, etc.) may not be satisfied fully. For example, d and e may be correlated with each other and d and e each may be serially correlated.

The purpose of our study is, therefore, to shed some light on the effect of measurement errors on the ordinary least squares estimator of β for the model given by (1.3), (1.4), and (1.6) for a large finite bivariate population when both variables are subject to correlated measurement errors for a large-scale sample survey situation.

In this paper, we present the following:

A. The mathematical development for the two-variable linear model [see equations (1.3) and (1.6)] to derive the bias factor of the ordinary least squares estimator of the parameter β, when both variables are subject to correlated measurement errors. This is presented in Section 2.

B. Estimators and estimates of the bias factor for some selected housing characteristics. This is presented in Section 3. We use two sets of data for calculation of the estimates. They are (1) a probability sample of about 5,000 housing units located in approximately 2,500 area segments of the United States. This sample was used for reinterview purposes by the Census Bureau after the 1960 Census of Housing in order to evaluate the accuracy of the statistics of housing characteristics. This project is known as the Content Evaluation Study (CES). The interviews were made in October 1960, six months after the census and the CES results were published in May, 1964. A detailed description of CES data is given by references [17]. (2) The second sample of housing units was drawn from six cities (Six-city data). This sample was chosen by the Census Bureau primarily for the purpose of evaluating the quality of housing conditions (e.g., sound, deteriorating, etc.) and methods of appraising the quality of housing conditions in 1964-65. Table 7 and reference [18] provide a detailed account of this sample.

C. Sensitivity analysis for the bias factor. This is also given in Section 3. For a set of hypothetical estimates of the parameters of the bias factor, sensitivity of the bias factor is examined.

2. THE MATHEMATICAL MODEL

We first define a set of survey conditions and assumptions for which the model is developed. Second, we derive the bias factor of the least squares estimator of the parameter β for the survey conditions and assumptions stated. Development of the model follows the work of Hansen, et.al.[9].

2.1 Survey Conditions

We use the term "survey" to mean census or sample surveys. This interpretation follows Hansen, et.al. [10].

A. Survey Conditions and Assumptions.

(1) We consider a large population of N elementary units, which is divided into M geographical areas (e.g., census tracts, enumeration districts, blocks, etc.). Each geographical area contains N_i elementary units, and thus

$$\sum_{i=1}^{M} N_i = N, \quad i=1, 2, \ldots, M.$$

(2) We postulate a simple random sample of n elementary units yielding n_i units from i-th area. Thus

$$\sum_{i=1}^{M} n_i = n$$

(3) We assume that $n_i \doteq \bar{n}$ for all i, where $\bar{n} = \frac{n}{M}$, the average sample size per geographical area.

(4) Each of M interviewers is assigned at random to one area and so every interviewer is responsible for $\bar{n}$ units.

(5) The process of collecting data by interviewers is conducted in such a way that measurement error is correlated within interviewer assignment areas, but is uncorrelated between the different interviewer assignment areas.[2/]

(6) We assume that the survey can be repeated independently under the same survey conditions.[3/]

(7) We assume that the ratios $\frac{n}{N}$ and $\frac{\bar{n}}{N_i}$ are small enough to ignore the finite population multipliers (i.e., $1-\frac{n}{N}$ and $1-\frac{\bar{n}}{N_i}$).

(8) We further assume that

$$\bar{n} - 1 \doteq \bar{n}$$

$$N_i - 1 \doteq N_i \text{ for all } i$$

2.2 Development of the Model

In this section, we introduce the definitions and notations first.

A. Definitions and Notations

Let x_{ijt}, y_{ijt} be the observed values of the variables x and y for the j-th sample unit of the i-th geographical area, when measurement is obtained at the t-th trial. The conditional expected values of x and y given the j-th unit of the i-th area are, say,

$$E_t(x_{ijt}|i,j) = X_{ij} \quad (1)$$

$$E_t(y_{ijt}|i,j) = Y_{ij} \quad (2)$$

where the expectation is taken over trials. Following Hansen,et. al. [10], we define the response deviation for x and y variables given the j-th sample unit of the i-th geographical area as follows:

$$d_{ijt} = x_{ijt} - X_{ij} \quad (3)$$

$$e_{ijt} = y_{ijt} - Y_{ij} \quad (4)$$

We assume that each of the error terms, d and e, follows a probability distribution[4/], and that the mean and variance of the distribution exist. Then, from equations (1) and (2), the conditional means, variances and covariance of d and e for a fixed j-th sample unit of the i-th geographical area as t varies are given by:

$$E_t(d_{ijt}|i,j) = 0 \quad (5)$$

$$E_t(e_{ijt}|i,j) = 0 \quad (6)$$

$$Var(d_{ijt}|i,j) = E_t(d^2_{ijt}|i,j) = \sigma^2_{d(ij)}, \text{ say} \quad (7)$$

$$Var(e_{ijt}|i,j) = E_t(e^2_{ijt}|i,j) = \sigma^2_{e(ij)}, \text{ say} \quad (8)$$

$$Cov(d_{ijt}e_{ijt}|i,j) = E_t(d_{ijt}e_{ijt}|i,j) = \sigma^2_{de(ij)}, \text{ say} \quad (9)$$

We further define the uncorrelated component of the response variance and covariance for a geographical area and for the entire population as follows[10].

$\frac{1}{N_i} \sum_j \sigma^2_{d(ij)} = \sigma^2_{d(i)}$, the simple variance for x for the i-th geographical area (10)

$\frac{1}{N_i} \sum_j \sigma^2_{e(ij)} = \sigma^2_{e(i)}$, the simple variance for y for the i-th geographical area (11)

$\frac{1}{N_i} \sum_j \sigma_{de(ij)} = \sigma_{de(i)}$, the simple response covariance for x and y for the i-th geographical area (12)

$\frac{1}{N} \sum_{i,j} \sigma^2_{d(ij)} = \sigma^2_e$, the simple response variance of the population for x (13)

$\frac{1}{N} \sum_{i,j} \sigma^2_{e(ij)} = \sigma^2_e$, the simple response variance of the population for y (14)

$\frac{1}{N} \sum_{i,j} \sigma_{de(ij)} = \sigma_{de}$, the simple response covariance of the population for x and y (15)

We next define the correlated component of the response variance and covariance in terms of the intra-class correlation coefficient $\delta_{d(i)}$ between d_{ijt} and $d_{ij't}$ for the i-th geographical area. The intraclass correlation coefficient of response deviations for the i-th geographical area. The intraclass correlation coefficient of response deviations for the i-th geographical area for x is defined by

$$\delta_{d(i)} = \frac{\sum_{j \neq}^{N_i} \sum_{j'}^{N_i} \sum_t E(d_{ijt} d_{ij't} | i,j,j')}{(N_i - 1) \sum_j^{N_i} \sum_t E\,(d^2_{ijt} | i,j)} \qquad (16)$$

From (7) and (10), we obtain,

$$\delta_{d(i)} = \frac{\sum_{j \neq}^{N_i} \sum_{j'}^{N_i} \sum_t E\,(d_{ijt} d_{ij't} | i,j,j')}{(N_i)(N_i - 1)\, \sigma^2_{d(i)}} \qquad (17)$$

hence

$$\delta_{d(i)} \sigma^2_{d(i)} = \frac{\sum_{j \neq}^{N_i} \sum_{j'}^{N_i} \sum_t E\,(d_{ijt} d_{ij't} | i,j,j')}{(N_i)(N_i - 1)} \qquad (18)$$

We call the quantity given by equation (18) the correlated component of the response variance of x for the i-th geographical area. Similarly, the correlated component of the response variance of y for the i-th geographical area is

$$\delta_{e(i)} \sigma^2_{e(i)} = \frac{\sum_{j \neq}^{N_i} \sum_{j'}^{N_i} \sum_t E\,(e_{ijt} e_{ij't} | i,j,j')}{(N_i)(N_i - 1)} \qquad (19)$$

The correlated component of the response covariance is defined similarly by defining the intra-class correlation of the response deviations for the i-th geographical area for x and y, i.e.,

$$\delta_{de(i)} = \frac{\sum_{j \neq}^{N_i} \sum_{j'}^{N_i} \sum_t E\,(d_{ijt} e_{ij't} | i,j,j')}{(N_i)(N_i - 1)\, \sigma_{d(i)} \sigma_{e(i)}} \qquad (20)$$

By multiplying both sides of (2) above by $\sigma_{de(i)}\, \rho^{-1}_{de(i)}$, we obtain the correlated component of the response covariance for the i-th geographical area:

$$\delta_{de(i)} \sigma_{de(i)} \sigma^{-1}_{de(i)} = \frac{\sum_{j \neq}^{N_i} \sum_{j'}^{N_i} \sum_t E(d_{ijt} e_{ij't} | i,j,j')}{(N_i)(N_i - 1)} \qquad (21)$$

where

$$\rho_{de} = \frac{\sigma_{de(i)}}{\sigma_{d(i)\ e(i)}} \qquad (22)$$

is the correlation coefficient of d, e, for the i-th geographical area.

The average correlated component of the response variance per geographical area is defined by

$$\overline{\sigma_d^2\delta_d} = \frac{1}{n}\sum_i^M n_i\ \sigma^2_{d(i)}\delta_{d(i)} \quad (23)$$

$$\doteq \frac{1}{M}\sum_i^M \sigma^2_{d(i)}\delta_{d(i)} \quad (24)$$

since we assumed that $n_i \doteq \bar{n}$ for all i.

Similarly, the average correlated component of the response covariance per geographical area is

$$\overline{\sigma_{de}\delta_{de}\rho^{-1}_{de}}$$

$$= \frac{1}{n}\sum_i^M n_i\ \sigma_{de(i)}\delta_{de(i)}\rho^{-1}_{de(i)} \quad (25)$$

$$\doteq \frac{1}{M}\sum_i \sigma_{de(i)}\delta_{de(i)}\rho^{-1}_{de(i)} \quad (26)$$

since $n_i \doteq \bar{n}$ for all i.

B. Bias of the Least Squares Estimator of β

As we stated earlier, our main objective is to study the effect of correlated measurement error on the ordinary least squares estimator in estimating β of the model given by (1.3), (1.4), and (1.6). For our survey situation we rewrite the model given by (1.3) and (1.6) by

$$Y_{ij} = \alpha + \beta X_{ij} \quad (27)$$

$$y_{ijt} = \alpha + \beta x_{ijt} + \omega_{ijt} \quad (28)$$

where α and β are parameters and the random variables X_{ij} and Y_{ij} are the conditional expected values of x_{ijt} and y_{ijt},

and $\omega_{ijt} = e_{ijt} - \beta d_{ijt}$ (29)

We observe, from a sample of n units, a set of values x_{ijt} and y_{ijt}; and estimate β using the ordinary least squares estimator b_t, which is defined by

$$b_t = \frac{\sum_i^M \sum_j^{N_i}(x_{ijt} - \bar{x}_t)(y_{ijt} - \bar{y}_t)}{\sum_i^M \sum_j^{N_i}(x_{ijt} - \bar{x}_t)^2}$$

$$= \frac{s_{xy(t)}}{s^2_{x(t)}} \quad (30)$$

where $\bar{x}_t$ and $\bar{y}_t$ are the sample means of x and y, respectively, for the t-th trial, i.e.,

$$\bar{x}_t = \frac{1}{n}\sum_i^M \sum_j^{n_i} x_{ijt} \quad (31)$$

$$\bar{y}_t = \frac{1}{n}\sum_i^M \sum_j^{n_i} y_{ijt} \quad (32)$$

And $s^2_{x(t)}$ and $s_{xy(t)}$ respectively are the sample variance and covariance for the t-th trial, i.e.,

$$s^2_{x(t)} = \frac{1}{n}\sum_i^M \sum_j^{n_i}(x_{ijt} - \bar{x}_t)^2 \quad (33)$$

$$s_{xy(t)}$$

$$= \frac{1}{n}\sum_i^M \sum_j^{n_i}(x_{ijt} - \bar{x}_t)(y_{ijt} - \bar{y}_t) \quad (34)$$

We are concerned with the bias of b_t. We derive the bias by taking expectation of b_t. The expectation is taken first over repeated trials for a fixed sample, and then over all possible samples. The ratio of expected values of the denominator and numerator of b_t is not necessarily equal to the expected value of b_t, since b_t is the ratio of two random variables. However, it is shown [4] that the differences between the expectation of the ratio and the ratio of expectation is small enough to ignore when the size of M is reasonably large for our survey conditions. Therefore, we evaluate the expectations of the denominator and numerator of b_t separately.

We find the expectation of the denominator of the b_t first.

$$\underset{st}{EE}\ s^2_{x(t)} = \underset{st}{EE}\ \frac{1}{n}\sum_i\sum_j (x_{ijt} - \bar{x}_t)^2$$

$$= \underset{st}{EE}\ \frac{1}{n}\sum_i\sum_j\left[(X_{ij}-\bar{X}) + (d_{ijt}-\bar{d}_t)\right]^2 \text{ by (3)} \tag{35}$$

where

$\bar{X} = \frac{1}{n}\sum_i^M\sum_j^{N_i} X_{ij}$, the sample mean of X_{ij}

$\bar{d}_t = \frac{1}{n}\sum_i^M\sum_j^{N_i} d_{ijt}$, the sample mean of d_{ijt} for the t-th trial.

For our Survey Condition, it can be shown that [4]

$$\underset{st}{EE}\ s^2_{x(t)} = \frac{n-1}{n}\left[\sigma^2_X+\sigma^2_d - \frac{1}{n(n-1)}\sum_i n_i\ (n_i-1)\sigma^2_{d(i)}\delta_{d(i)}\right] \tag{36}$$

$$\doteq \sigma^2_X+\sigma^2_d - \frac{1}{M^2}\sum_i \sigma^2{}_{d(i)}\delta_{d(i)}$$

$$= \sigma^2_X+\sigma^2_d - \frac{1}{M}\overline{\sigma^2_d\delta_d}$$

since we assumed that $n-1 \doteq \bar{n}$, and

$\frac{1}{M}\sum \sigma^2_{d(i)}\delta_{d(i)} = \overline{\sigma^2_d\delta_d}$ by equation (24);

and where

$\sigma^2_X = \frac{1}{M}\sum_i^M\sum_j^{N_i}(X_{ij}-\bar{X}_{(p)})^2$, the variance of X_{ij}, (38)

$\bar{X}_{(p)} = \frac{1}{N}\sum_i^M\sum_j^{N_i} X_{ij}$, the population mean of X_{ij}, (39)

and σ^2_d and $\sigma^2_{d(i)}\delta_{d(i)}$ are defined by (13) and (18) respectively.

In a similar manner, we can show the expected value of the numerator of the b_t to be

$$\underset{st}{EE}\ s_{xy(t)} = \frac{n-1}{n}\left[\sigma_{XY}+\sigma_{de} - \frac{1}{n(n-1)}\sum_i n_i(n_i-1)\sigma_{de(i)}\delta_{de(i)}\rho^{-1}_{de(i)}\right]$$

$$\doteq \sigma_{XY}+\sigma_{de} - \frac{1}{M}\overline{\sigma_{de}\delta_{de}\rho^{-1}_{de}} \tag{40}$$

where

$$\sigma_{XY} = \frac{1}{N}\sum_i^M\sum_j^{N_i}(X_{ij}-\bar{X}_{(p)})(Y_{ij}-\bar{Y}_{(p)}) \tag{41}$$

the covariance of X_{ij} and Y_{ij}; and σ_{de}, $\sigma_{de(i)}\delta_{de(i)}\rho^{-1}_{de(i)}$, and $\overline{\sigma_{de}\delta_{de}\rho^{-1}_{de}}$ are given by (15), (21), and (26) respectively.

We denote the ratio of expected values of the numerator and denominator of b_t by β^*. Thus,

$$\beta^* = \frac{\underset{st}{EE}\ s_{xy(t)}}{\underset{st}{EE}\ s^2_{x(t)}} \doteq \frac{\sigma_{XY}+\sigma_{de} - \frac{1}{M}\overline{\sigma_{de}\delta_{de}\rho^{-1}_{de}}}{\sigma^2_X+\sigma^2_d - \frac{1}{M}\overline{\sigma^2_d\delta_d}} \tag{42}$$

For the sake of simplicity in writing, we define

$$\sigma^2_{d(T)} = \sigma^2_d - \frac{1}{M}\sum_i \overline{\sigma^2_d\delta_d} \tag{43}$$

$$\sigma_{de(T)} = \sigma_{de} - \frac{1}{M}\sum_i \overline{\sigma_{de}\delta_{de}\rho^{-1}_{de}} \tag{44}$$

$$\sigma^2_{x(T)} = \sigma^2_X + \sigma^2_{d(T)} \tag{45}$$

$$\sigma_{xy(T)} = \sigma_{XY} + \sigma_{de(T)} \tag{46}$$

Using (43), (44), (45), and (46) above, we write

$$\beta^* = \frac{\sigma_{XY} + \sigma_{de(T)}}{\sigma^2_X + \sigma^2_{d(T)}} = \frac{\sigma_{XY(T)}}{\sigma^2_{X(T)}} \tag{47}$$

Factoring $\beta = \frac{\sigma_{XY}}{\sigma^2_X}$ out from (47) above, we have

$$\beta^* = (\beta)\,\frac{(1+\sigma_{de(T)}/\sigma_{XY})}{(1+\sigma^2_{d(T)}/\sigma^2_X)} \qquad (48)$$

The second factor of the right hand member of (48) above is defined to be the component bias factor of the b_t for our Survey Conditions. We denote it by E_1, i.e.,

$$E_1 = \frac{1+\sigma_{de(T)}/\sigma_{XY}}{1+\sigma^2_{d(T)}/\sigma^2_X} \qquad (49)$$

Hence, we have

$$\beta^* = (\beta)(E_1) \qquad (50)$$

or

$$\frac{\beta^*}{\beta} = E_1 \qquad (51)$$

or

$$\frac{(\beta^*-\beta)}{\beta} = (E_1-1) \qquad (52)$$

Assuming that M is so large that $\underset{st}{EE}\, b_t - \beta^* \doteq 0$ we can note from (49),(50), and (52) that $E_1 > 1$ indicates an over-estimation of β by b_t on the average; $E_1 < 1$ means an underestimation of β; and that when $E_1 = 1$, b_t is unbiased of β. And the bias factor E_1 is a function of uncorrelated components of response variance and covariance and the correlated components of response variance and covariance. In the following section, we estimate the bias factor E_1 from the two sets of sample data which we described earlier.

3. ESTIMATORS AND ESTIMATES

In this section, we discuss estimators and estimates of σ^2_d, σ_{de}, $\overline{\sigma^2_d\delta_d}$, and $\overline{\sigma_{de}\delta_{de}\rho^{-1}_{de}}$, first; and then we discuss estimators and estimates of $\sigma^2_{x(T)}$, $\sigma_{xy(T)}$, σ^2_d/σ^2_X, and σ_{de}/σ_{XY}; and finally we discuss the estimates of E_1.

3.1 The Estimators and Estimates of σ^2_d and σ_{de}.

As an estimator of σ^2_d Hansen, et.al. [9]more or less give

$$\frac{g}{2} = \frac{1}{2n}\sum_i^M \sum_j^{n_i}(x_{ijtG} - x_{ijt'G'})^2 \qquad (53)$$

where $g = \frac{1}{n}\sum_i \sum_j (x_{ijtG} - x_{ijt'G'})$ stands for "gross difference rate" [2,0], and t and t' respectively refer to t-th and t'th trials and G and G' respectively refer to G-th and G'th survey conditions. Following Hansen, et.al. [9] we can show that

$$\underset{st}{EE}\,\frac{g}{2} = \sigma^2_d \qquad (54)$$

if $\underset{s}{E}(x_{ijtG}) = \underset{s}{E}(x_{ijt'G'})$ and if repeated surveys are done independently so that the trial to trial covariance is zero. According to Bailer [2], the between-trial covariance is relatively small for the items she studied for a reinterview procedure for which the interviewers did not have access to the Census Data (original data) and reconciliation was not made after reinterview. The CES data we used in our study is obtained by the same interview procedure as the one just mentioned above, although the items she studied are not the same as the ones we studied. As for estimates of the bias $\underset{s}{E}(x_{ijtG} - x_{ijt'G})$, Bailer's study [2] did not seem to show any definite conclusion on the differences in estimates of the bias for different interview procedures. However, Bailer points out that, for a large sample, "a reinterview procedure which specifies that the reinterview be closer in time to the original interview" [P.60, Ref. 2] than the CES data (six months lag between original interview and reinterview) seem to have smaller bias.

In short, we are not sure about the magnitudes of the between-trial covariance and bias due to different survey procedures for the housing items included in our study. But the estimation of these terms are beyond the scope of our study. Instead, we assert that the assumptions and survey conditions stated at the outset hold so that $\frac{g}{2}$ is a good estimator of σ^2_d. Pritzker [16] (see also [2]) gives an estimator of σ_{de} by

$$\frac{h}{2} = \frac{1}{2n} \sum_{i}^{M} \sum_{j}^{n_i} (x_{ijtG} - x_{ijt'G'})(y_{ijtG} - y_{ijt'G'}) \quad (55)$$

where

$$h = \frac{1}{n} \sum_{i} \sum_{j} (x_{ijtG} - x_{ijt'G'})(y_{ijtG} - y_{ijt'G'})$$

Following Pritzker [16], we can show that

$$\underset{st}{EE} \frac{h}{2} = \sigma_{de} \quad (56)$$

if $\underset{s}{E}(x_{ijtG}) = \underset{s}{E}(x_{ijt'G'})$ and $\underset{s}{E}(y_{ijtG}) = \underset{s}{E}(y_{ijt'G'})$ and if independent repetitions of a survey are made.

Again, we say that the survey conditions we assumed hold and $\frac{h}{2}$ is a good estimator of σ_{de}.

The sample stimates of σ_d^2 for some selected housing characteristics are given in Table 3. The sample estimates of σ_{de} for some selected housing characteristics are also calculated from the two sets (Census results and CES Data) of sample data in 1967 for the first time. These estimates are given in Table 4.

3.2 Estimators and Estimates of $\overline{\sigma_d^2 \delta_d}$ and $\overline{\sigma_{de}\delta_{de}\rho_{de}^{-1}}$.

The Response Variance Study [1] conducted by Bailer at the Census Bureau shows estimates of "ρ_d, the interclass correlation between response deviations of different units assigned to the same interviewer" (which is comparable to δ_d in our study) for population and housing characteristics. The estimates in that study were made on the basis of an interpenetrated sample design using an estimator similar to the one given below:

$$\frac{1}{\bar{n}} (s_{xbt}^2 - s_{xwt}^2) \quad (56)$$

where

$$s_{xbt}^2 = \frac{1}{M-1} \left[\bar{n} \sum (\bar{x}_{it} - \bar{x}_t)^2 \right] \text{, between-interviewers variance} \quad (57)$$

$$\bar{x}_{it} = \frac{1}{\bar{n}} \sum_{j}^{\bar{n}} x_{ijt} \text{, the sample mean of x for the i-th geographical area}$$

$$s_{xwt}^2 = \frac{1}{M(\bar{n}-1)} \sum_{ij}^{M\bar{n}} (x_{ijt} - \bar{x}_{it})^2 \text{, within-interviewers variance} \quad (58)$$

According to the response variance study by Bailer of the Census Bureau [1] the magnitudes of the estimates of the average correlated response variance decreases as the interviewer assignment areas increase. In fact, the study concludes:

> "Though the rate of decrease is not constant over all population and housing characteristics, it is reasonable to assume that the relationship between ρ_d and the size of the assignment area is described by an exponential function, ..." (pp.3-4, [1]).

The Census Bureau study [1] did not estimate δ_{de}, the intra-class correlation for d and e. Furthermore, the estimates of the housing characteristics in which we are interested in our study were not included in the Census Bureau study. Therefore, we estimated the average correlated component of response variance and covariance for some selected housing characteristics using the Six-City Data.

For estimating $\overline{\sigma_d^2 \delta_d}$ we used the following estimator:

$$\frac{\bar{n}}{\bar{n}-1} (s_{\bar{x}t}^2 - s_{xt}^2/\bar{n}) \quad (59)$$

and for estimating $\overline{\sigma_{de}\delta_{de}\rho_{de}^{-1}}$ we used

$$\frac{\bar{n}}{\bar{n}-1} (s_{\bar{x}\bar{y}t} - s_{xyt}/\bar{n}) \quad (60)$$

where

$$s_{\bar{x}t}^2 = \frac{\bar{n}}{n} \sum_{i}^{M} (\bar{x}_{it} - \bar{x}_t)^2 \text{, sample between-area variance} \quad (61)$$

$$s_{\bar{x}\bar{y}t} = \frac{\bar{n}}{n} \sum_{i}^{M} (\bar{x}_{it} - \bar{x}_t)(\bar{y}_{it} - \bar{y}_t) \text{, sample between-area covariance} \quad (62)$$

and s_{xt}^2 and s_{xyt} are given by (33) and (34).

The estimator given by (56) is the same as the estimator given by (59) if M-1=M. And also, it can be shown that the estimators given by (56) and (59) are unbiased estimators of $\overline{\sigma_d^2\delta_d}$ if our assumptions and survey conditions hold. Similarly, the estimator (60) is an unbiased estimator of $\overline{\sigma_{de}\delta_{de}\rho_{de}^{-1}}$.

The estimates of $\overline{\sigma_d^2\delta_d}$ and $\overline{\sigma_{de}\delta_{de}\rho_{de}^{-1}}$ are given by Tables 1A, 1B, and 1C.

3.3 Estimates of $\sigma_{d(T)}^2$ and $\sigma_{de(T)}$.

Tables 2A, 2B, and 2C show the sample estimates of σ_d^2, σ_{de}, $\frac{1}{M}\overline{\sigma_{de}\delta_{de}\rho_{de}^{-1}}$. From these estimates we note that, although $\sigma_{d(T)}^2 = \sigma_d^2 - \frac{1}{M}\overline{\sigma_d^2\delta_d}$ and $\sigma_{de(T)} = \sigma_{de} - \frac{1}{M}\overline{\sigma_{de}\delta_{de}\rho_{de}^{-1}}$, in most cases $\frac{1}{M}\overline{\sigma_d^2\delta_d}$ and $\frac{1}{M}\overline{\sigma_{de}\delta_{de}\rho_{de}^{-1}}$ are negligible compared with σ_d^2 and σ_{de}. For example, almost all of the estimates of $\frac{1}{M}\overline{\sigma_d^2\delta_d}$ for block-sized areas is zero (see Table 2A) and the largest value for the ratio of $\frac{1}{M}\overline{\sigma_d^2\delta_d}$ to σ_d^2 is only about 0.1 (see Table 2A, Shreveport, "Bath for exclusive use", tract size). The largest value for the ratio of $\frac{1}{M}\overline{\sigma_{de}\delta_{de}\rho_{de}^{-1}}$ to σ_{de}, however, is about 0.3 (see Table 2C, Shreveport, "Owner occupied units," tract size). Nevertheless, most of the estimates given in Tables 2B and 2C indicate that $\frac{1}{M}\overline{\sigma_{de}\delta_{de}\rho_{de}^{-1}}/\sigma_{de}$ is very small.

Therefore, at least from these estimates, we may conclude that

$$\sigma_{d(T)}^2 \doteq \sigma_d^2 \tag{63}$$

$$\sigma_{de(T)} \doteq \sigma_{de} \tag{64}$$

Hence, from (45), (46), and (49), we have

$$\sigma_{x(T)}^2 \doteq \sigma_X^2 + \sigma_d^2 \tag{65}$$

$$\sigma_{xy(T)} \doteq \sigma_{XY} + \sigma_{de} \tag{66}$$

$$E_1 \doteq \frac{1+\sigma_{de}/\sigma_{XY}}{1+\sigma_d^2/\sigma_X^2} \tag{67}$$

Let the sample variance of x for the t-th trial be denoted by s_{xt}^2 and that for the t'th trial by $s_{xt'}^2$, and let the average of the two variances be denoted by $s_{x(T)}^2$, i.e., $s_{x(T)}^2 = \frac{1}{2}(s_{xt}^2 + s_{xt'}^2)$ (68)

$s_{x(T)}^2$ is the estimator of $\sigma_{x(T)}^2$.5/ Hence, $g/2s_{x(T)}^2$ will estimate $\sigma_d^2/\sigma_{x(T)}^2$.

Assuming that $\frac{g}{2}$ is an unbiased estimator of σ_d^2 and that $\sigma_{x(T)}^2 \doteq \sigma_d^2+\sigma_X^2$ we can use, as an estimator of $\sigma_X^2/\sigma_{x(T)}^2$, $1 - \frac{g}{2s_{x(T)}^2}$. And furthermore, as (69)

an estimator of σ_d^2/σ_X, we use

$$\left(\frac{g}{2s_{x(T)}^2}\right)\left(1 - \frac{g}{2s_{x(T)}^2}\right)^{-1} \tag{70}$$

The estimates of σ_{de}/σ_{XY} are similarly obtained. The sample estimates of σ_d^2/σ_x^2 and σ_{de}/σ_{xy} are given in Tables 3 and 4.

Table 3 shows that the range of the estimates of σ_d^2/σ_x^2 for the given housing characteristics is .1099 to 3.0388.

Assuming that (63) and (64) hold, (i.e., $\frac{1}{M}\overline{\sigma_d^2\delta_d}$ and $\frac{1}{M}\sigma_{de}\delta_{de}\rho_{de}^{-1}$ are negligible), we can see that

$$0 \le \frac{\sigma_d^2}{\sigma_{x(T)}^2} \le 1 \tag{71}$$

and

$$0 \le \frac{\sigma_d^2}{\sigma_X^2} \le \infty \tag{72}$$

In other words, the larger the estimates of $\sigma_d^2/\sigma_{x(T)}^2$, the larger the estimates of σ_d^2/σ_X^2. In fact, if an estimate of $\sigma_d^2/\sigma_{x(T)}^2$ is close to 1 (meaning a highly inaccurate measurement) a corresponding value of σ_d^2/σ_X^2 is very large.

Similarly, we can see that

$$- \infty \le \frac{\sigma_{de}}{\sigma_{xy(T)}} \le \infty \qquad (73)$$

and

$$- 1 \le \frac{\sigma_{de}}{\sigma_{XY}} \le \infty \qquad (74)$$

since σ_{de} and σ_{xy} do not necessarily take the same sign. The actual estimates of $\sigma_{de}/\sigma_{xy(T)}$ for the selected housing characteristics shown in Table 4 range from -.1106 to .5493 and the estimates of σ_{de}/σ_{XY} vary from -.0996 to 1.2188.

3.4 Estimates of E_1

We first discuss direct estimates of E_1, which are calculated from the estimates of σ_d^2/σ_X^2 in Table 3 and σ_{de}/σ_{XY} in Table 4, and then show the sensitivity of E_1.

(1) Direct estimates of E_1.

Table 5 shows the estimates of E_1 along with estimates of β^* and β. As seen in Table 5, we note that more than half of the estimates of E_1 are greater than one. In other words, for more than half of the pairs of housing variables in Table 5, σ_{de}/σ_{XY} is greater than σ_d^2/σ_X^2 in magnitude. This means that an important contribution is made to the effect of measurement error on the estimator of β by the simple response covariance, σ_{de}. It is wrong, therefore, to assume in all cases that $\sigma_{de} = 0$ and say that the effect upon the least squares estimator of β due to measurement error is "attenuating."

It is interesting to note that there is a certain consistency in the variation of E_1 over different variables. First, we notice that the value of E_1 is greater than one for the variables, "owner occupied units" and "units with bath for exclusive use," in all three dependent variables ("sound units," "deteriorating units," and "dilapidated units"). This leaves the value of E_1 for the other two variables ("renter occupied units" and "units with shared or no bath") to be less than one.

The magnitudes of E_1 range from .5680 to 1.5164, and the effect of measurement error on the estimator of β for the housing characteristics shown in Table 5 is clearly seen from the estimates of E_1 or from the comparison of the estimates β^* and β. We can see that the attenuation of β was largest for the pair of variables, "sound units" vs. "units with shared or no bath" (i.e., the estimate of β is -.9602 whereas the estimate of β^* is -.5454) and the magnitude of the overestimates of β was greatest for the pair of variables, "sound units" vs. "units with bath for exclusive use" (the estimate of β is .2194 and the estimate of β^* is .3327). For the pair of variables, "deteriorating units" vs. "renter occupied",there is almost no effect of measurement error on the estimator of β.

(2) Sensitivity analysis of the estimates of E_1.

So far we have investigated the sample estimates of E_1 for some housing characteristics. In this section, we will study the sensitivity of the magnitudes of E_1 for different values of $\sigma_d^2/\sigma_{x(T)}^2$ and $\sigma_{de}/\sigma_{xy(T)}$ based on the largest magnitude of the sample estimates given in Tables 3 and 4.

From what we have observed in Tables 3 and 4, .8 is the largest estimate for $\sigma_d^2/\sigma_{x(T)}^2$ and $\sigma_{de}/\sigma_{xy(T)}$. Therefore, we take 25, 50, 75, and 100 percent of .8 and add to these different values -.2 for the smallest value of $\sigma_{de}/\sigma_{xy(T)}$ and zero for $\sigma_d^2/\sigma_{x(T)}^2$. And then we compute different values of E_1 for these different values of $\sigma_d^2/\sigma_{x(T)}^2$ and $\sigma_{de}/\sigma_{xy(T)}$. The results of these computations are shown in Table 6 and Figure 1.

Examining the sensitivity of the values of E_1, we can clearly see from Figure 1 and Table 6 that, for a given value of $\sigma_d^2/\sigma_{x(T)}^2$ the values of E_1 increase with increasing rate as the values of $\sigma_{de}/\sigma_{xy(T)}$ increase by .2 from -.2 to .8. We also notice from Figure 1 that the rate of increase E_1 over the different values of $\sigma_{de}/\sigma_{xy(T)}$ gets smaller as the value of $\sigma_d^2/\sigma_{x(T)}^2$ increases. On the other hand, the rate of increase in the values of E_1 for the different values of $\sigma_d^2/\sigma_{x(T)}^2$ in Figure 1 is smaller for $\sigma_{de}/\sigma_{xy(T)} \leq .4$ than for $\sigma_{de}/\sigma_{xy(T)} > .4$.

Furthermore, we can see from Figure 1 and Table 6 that the values of E_1 will be greater than one (i.e., $E_1>1$) when $\sigma_{de}/\sigma_{xy(T)} > \sigma_d^2/\sigma_{x(T)}^2$, and $E_1 < 1$ when $\sigma_{de}/\sigma_{xy(T)}<\sigma_d^2/\sigma_{x(T)}^2$. And of course, $E_1=1$ when $\sigma_{de}/\sigma_{xy(T)}=\sigma_d^2/\sigma_{x(T)}^2$, and $E_1<1$ when $\sigma_{de}/\sigma_{xy(T)}<0$ for all values of $\sigma_d^2/\sigma_{x(T)}^2$.

As we noted earlier, $E_2<1$ means underestimation of β, $E_1>1$ means overestimation of β, and $E_1=1$ means that β is neither underestimated nor overestimated.

4. SUMMARY

For the survey conditions assumed and the variables investigated in our study, we find that the major contribution to the response variances and covariance is from the uncorrelated components of response variances and covariance.

The sample estimates of E_1 for some housing characteristics given in Table 5 reveal that we estimate that the mean value of b_t underestimates β by as much as 43 percent and overestimates β by as much as 57 percent.

Table 6 and Figure 1 show the sensitivity of the values of E_1. As seen in Figure 1 the values of E_1 are more sensitive to $\sigma_{de}/\sigma_{xy(T)} \geq .5$ and $0 \leq \sigma_d^2/\sigma_{x(T)}^2 \leq .6$ than to $-.2 \leq \sigma_{de}/\sigma_{xy(T)} < .5$ and $.6 < \sigma_d^2/\sigma_{x(T)}^2 < 1$.

FOOTNOTES

1/ There are situations in which this exact functional relation may not hold. For example, for some cases, the linear model with disturbance term ε, i.e., $Y=\alpha+\beta X+\varepsilon$. But since the model (1.3) is a fundamental one and often discussed, our study is based on this model.

2/ Of course, if the interviewers are supervised by different persons, then the measurement error between different interviewer assignment areas may be correlated.

3/ Independent repetitions of a survey under constant survey conditions can hardly be achieved in practice, but the postulate given here will be the basis for defining the variance and covariance of the measurement errors. See Reference [2] for a case of a dependent reinterview situation.

4/ We do not assume any particular form of probability distribution for d and e.

5/ The Census Bureau used an estimator similar to $s_{x(T)}^2$ (i.e., $\frac{1}{2}(p_1q_1+p_2q_2)$, where p_iq_i is the sample variance for "0,1" variable for sample of size one for the i-th trial. See [17] for detail.

Table 1A.--Estimates of the Average Correlated Component of Response Variance for Some Housing Variables.

City and variable	Estimates of $\overline{\sigma_d^2 \delta_d}$			City and variable	Estimates of $\overline{\sigma_d^2 \delta_d}$		
	Block	ED	Tract		Block	ED	Tract
Camden, New Jersey:				Louisville, Kentucky:			
Owner occupied units	.0694	.0498	.0238	Owner occupied units	.0824	.0764	.0632
Deteriorating units .	.0428	.0328	.0207	Deteriorating units .	.0353	.0326	.0192
Dilapidated units . .	.0104	.0075	.0038	Dilapidated units . .	.0092	.0075	.0037
Substandard units . .	.0230	.0178	.0138	Substandard units . .	.0371	.0366	.0286
Bath for exclusive use	.0204	.0132	.0108	Bath for exclusive use	.0351	.0320	.0252
Units built 1939 or earlier.	.0958	.0665	.0373	Units built 1939 or earlier.	.1331	.1200	.0982
Monthly rent less than $80.	.0395	.0259	.0177	Monthly rent less than $80.	.1047	.0993	.0867
Cleveland, Ohio:				Shreveport, Louisiana:			
Owner occupied units	.0672	.0623	.0526	Owner occupied units	.0776	.0748	.0632
Deteriorating units .	.0415	.0379	.0188	Deteriorating units .	.0391	.0371	.0196
Dilapidated units . .	.0092	.0084	.0046	Dilapidated units . .	.0178	.0154	.0048
Substandard units . .	.0241	.0226	.0159	Substandard units . .	.0764	.0688	.0550
Bath for exclusive use	.0157	.0137	.0096	Bath for exclusive use	.0738	.0634	.0514
Units built 1939 or earlier.	.0737	.0681	.0547	Units built 1939 or earlier.	.1265	.1135	.0810
Monthly rent less than $80.	.0672	.0632	.0532	Monthly rent less than $80.	.1139	.1076	.0899
Fort Wayne, Indiana:				South Bend, Indiana:			
Owner occupied units	.0501	.0463	.0375	Owner occupied units	.0523	.0485	.0342
Deteriorating units	.0290	.0207	.0102	Deteriorating units	.0192	.0154	.0072
Dilapidated units . .	.0044	.0027	.0011	Dilapidated units . .	.0050	.0018	.0006
Substandard units . .	.0105	.0095	.0059	Substandard units . .	.0176	.0118	.0079
Bath for exclusive use	.0115	.0067	.0042	Bath for exclusive use	.0184	.0092	.0067
Units built 1939 or earlier.	.1282	.1149	.0969	Units built 1939 or earlier.	.1536	.1385	.1180
Monthly rent less than $80.	.0889	.0774	.0634	Monthly rent less than $80.	.0891	.0789	.0661

Source: Six-city data, Bureau of the Census
(See Table 7 of this article)

Table 1B.--Estimates of the Average Correlated Component of Response Covariance for Some Housing Variables.

("Units Deteriorating" as Dependent Variables.)

City and variable	Estimates of $\sigma_{de}\delta_{de}\rho_{de}^{-1}$			City and variable	Estimates of $\sigma_{de}\delta_{de}\rho_{de}^{-1}$		
	Block	ED	Tract		Block	ED	Tract
Camden, New Jersey:				Shreveport, Louisiana:			
Owner occupied units	-.0094	-.0090	-.0139	Owner occupied units	-.0288	-.0258	-.0218
Substandard units .	.0143	.0123	.0123	Substandard units .	.0390	.0341	.0283
Bath for exclusive use.	-.0125	-.0099	-.0106	Bath for exclusive use.	-.0361	-.0319	-.0267
Units built 1939 or earlier.	.0228	.0170	.0136	Units built 1939 or earlier.	.0258	.0242	.0171
				Monthly rent less than $80	.0457	.0414	.0348
Cleveland, Ohio:							
Owner occupied units	-.0208	-.0197	-.0189				
				South Bend, Indiana:			
Bath for exclusive use.	-.0099	-.0097	-.0077	Owner occupied units	-.0150	-.0133	-.0113
Units built 1939 or earlier.	.0147	.0141	.0128	Bath for exclusive use.	-.0084	-.0065	-.0053
Monthly rent less than $80	.0248	.0243	.0224	Monthly rent less than $80	.0244	.0201	.0166
Fort Wayne, Indiana:				Louisville, Kentucky:			
Owner occupied units	-.0190	-.0149	-.0125	Owner occupied units	-.0176	-.0186	-.0192
Substandard units	.0100	.0090	.0058	Substandard units	.0239	.0229	.0181
Bath for exclusive use.	-.0051	-.0054	-.0041	Bath for exclusive use.	-.0213	-.0205	-.0164
Units built 1939 or earlier.	.0208	.0200	.0172	Units built 1939 or earlier.	.0308	.0294	.0250
Monthly rent less than $80	.0308	.0251	.0220	Monthly rent less than $80	.0324	.0210	.0143

Source: Six-city data, Bureau of the Census
(See Table 7 of this article)

Table 1C.--Estimates of the Average Correlated Component of Response Covariance for Some Housing Variables.

("Units Dilapidated" as Dependent Variable.)

City and variable	Estimates of $\overline{\sigma_{de}\delta_{de}\rho_{de}^{-1}}$		
	Block	ED	Tract
Camden, New Jersey:			
Owner occupied units	-.0032	-.0045	-.0038
Substandard units .	.0121	.0090	.0058
Bath for exclusive use.	-.0082	-.0067	-.0046
Cleveland, Ohio:			
Owner occupied units	-.0061	-.0061	-.0059
Substandard units .	.0103	.0096	.0061
Bath for exclusive use.	-.0028	-.0028	-.0025
Units built 1939 or earlier.	.0035	.0033	.0028
Monthly rent less than $80	.0074	.0072	.0072
Fort Wayne, Indiana:			
Owner occupied units	-.0046	-.0041	-.0031
Substandard units .	.0056	.0036	.0018
Bath for exclusive use.	-.0036	-.0020	-.0010
Monthly rent less than $80	.0068	.0064	.0054

City and variable	Estimates of $\overline{\sigma_{de}\delta_{de}\rho_{de}^{-1}}$		
	Block	ED	Tract
Louisville, Kentucky:			
Owner occupied units	-.0073	-.0074	-.0076
Substandard units .	.0124	.0113	.0080
Bath for exclusive use.	-.0093	-.0088	-.0069
Units built 1939 or earlier.	.0087	.0084	.0072
Monthly rent less than $80	.0108	.0103	.0098
Shreveport, Indiana:			
Owner occupied units	-.0118	-.0112	-.0071
Substandard units .	.0249	.0206	.0134
Bath for exclusive use.	-.0207	-.0171	-.0123
Units built 1939 or earlier.	.0104	.0085	.0043
Monthly rent less than $80	.0197	.0170	.0134
South Bend, Indiana:			
Owner occupied units	-.0037	-.0030	-.0023
Substandard units .	.0060	.0034	.0017

Source: Six-city data, Bureau of the Census
(See Table 7 of this article)

Table 2A.--Estimates of the Simple Response Variance and the Average Correlated Component of Response Variance Divided by M.

Numbers in the parentheses show the approximate number of geographical areas (i.e., M).*

City and housing characteristics	Estimates of			
	σ_d^2 1/	$\frac{1}{M}\overline{\sigma_d^2\delta_d}$ 2/		
		Block	ED	Tract
Camden, New Jersey:		(1,100)	(100)	(30)
Owner occupied units. . . .	.0230	.0001	.0005	.0008
Bath for exclusive use. . .	.0124	.0000	.0001	.0001
Deteriorating	.0872	.0000	.0003	.0007
Dilapidated	.0298	.0000	.0001	.0004
Cleveland, Ohio:		(4,400)	(1,000)	(200)
Owner occupied units. . . .	.0230	.0000	.0001	.0003
Bath for exclusive use. . .	.0124	.0000	.0000	.0001
Deteriorating	.0872	.0000	.0000	.0001
Dilapidated	.0298	.0000	.0000	.0000
Fort Wayne, Indiana:		(2,100)	(200)	(40)
Owner occupied units. . . .	.0230	.0000	.0003	.0009
Bath for exclusive use. . .	.0124	.0000	.0000	.0001
Deteriorating	.0872	.0000	.0001	.0003
Dilapidated	.0298	.0000	.0000	.0000
Louisville, Kentucky:		(3,000)	(500)	(100)
Owner occupied units. . . .	.0230	.0000	.0002	.0006
Bath for exclusive use. . .	.0124	.0000	.0001	.0003
Deteriorating	.0872	.0000	.0001	.0002
Dilapidated	.0298	.0000	.0000	.0000
Shreveport, Louisiana:		(2,000)	(200)	(40)
Owner occupied units. . . .	.0230	.0000	.0004	.0016
Bath for exclusive use. . .	.0124	.0000	.0003	.0013
Deteriorating	.0872	.0000	.0002	.0005
Dilapidated	.0298	.0000	.0001	.0001
South Bend, Indiana:		(1,700)	(200)	(30)
Owner occupied units. . . .	.0230	.0000	.0002	.0011
Bath for exclusive use. . .	.0124	.0000	.0000	.0002
Deteriorating	.0872	.0000	.0000	.0002
Dilapidated	.0298	.0000	.0000	.0000

* See Table 7 for the values of M.
1/ The estimates shown in this table are transcribed from Table 3.
2/ See Table 1A for the estimates of $\overline{\sigma_d^2\delta_d}$.

Table 2B.--Estimates of the Simple Response Covariance and the Average Correlated Response Covariance Divided by M.

("Units Deteriorating" as Dependent Variable.)

Numbers in the parentheses show the approximate number of geographical areas (i.e., M).*

City and housing characteristics	σ_{de} 1/	Estimates of $\frac{1}{M}\,\overline{\sigma_{de}\delta_{de}\rho_{de}^{-1}}$ 2/		
		Block	ED	Tract
Camden, New Jersey:		(1,100)	(100)	(30)
Owner occupied units.	-.0027	.0000	-.0001	-.0005
Bath for exclusive use. . . .	-.0058	.0000	-.0001	-.0004
Cleveland, Ohio:		(4,400)	(1,000)	(200)
Owner occupied units.	-.0027	.0000	.0000	-.0001
Bath for exclusive use. . . .	-.0058	.0000	.0000	.0000
Fort Wayne, Indiana:		(2,100)	(200)	(40)
Owner occupied units.	-.0027	.0000	-.0001	-.0003
Bath for exclusive use. . . .	-.0058	.0000	.0000	-.0001
Louisville, Kentucky:		(3,000)	(500)	(100)
Owner occupied units.	-.0027	.0000	.0000	-.0002
Bath for exclusive use. . . .	-.0058	.0000	.0000	-.0002
Shreveport, Louisiana:		(2,000)	(200)	(40)
Owner occupied units.	-.0027	.0000	-.0001	-.0005
Bath for exclusive use. . . .	-.0058	.0000	-.0002	-.0006
South Bend, Indiana:		(1,700)	(200)	(30)
Owner occupied units.	-.0027	.0000	-.0001	-.0003
Bath for exclusive use. . . .	-.0058	.0000	.0000	-.0002

* See Table 7 for the values of M.
1/ The estimates shown in this table are transcribed from Table 4.
2/ See Table 1B for the estimates of $\overline{\sigma_{de}\delta_{de}\rho_{de}^{-1}}$.

Table 2C.--Estimates of the Simple Response Covariance and the Average Correlated Response Covariance Divided by M.

("Units Dilapidated" as Dependent Variable.)

Numbers in the parentheses show the approximate number of geographical areas (i.e., M).*

City and housing characteristics	Estimates of			
	σ_{de} 1/	$\frac{1}{M}\ \overline{\sigma_{de}\delta_{de}\rho_{de}^{-1}}$ 2/		
		Block	ED	Tract
Camden, New Jersey:		(1,100)	(100)	(30)
Owner occupied units.	.0007	.0000	-.0000	-.0001
Bath for exclusive use. . . .	-.0045	.0000	-.0001	-.0002
Cleveland, Ohio:		(4,400)	(1,000)	(200)
Owner occupied units.	.0007	.0000	.0000	.0000
Bath for exclusive use. . . .	-.0045	.0000	.0000	.0000
Fort Wayne, Indiana:		(2,100)	(200)	(40)
Owner occupied units.	.0007	.0000	0000	-.0001
Bath for exclusive use. . . .	-.0045	.0000	.0000	.0000
Louisville, Kentucky:		(3,000)	(500)	(100)
Owner occupied units.	.0007	.0000	.0000	-.0001
Bath for exclusive use. . . .	-.0045	.0000	.0000	-.0001
Shreveport, Louisiana:		(2,000)	(200)	(40)
Owner occupied units.	.0007	.0000	-.0001	-.0002
Bath for exclusive use. . . .	-.0045	.0000	-.0001	-.0003
South Bend, Indiana:		(1,700)	(200)	(30)
Owner occupied units.	.0007	.0000	.0000	-.0001
Bath for exclusive use. . . .	-.0045	--	--	--

*See Table 7 for the values of M.

1/ The estimates shown in this table are transcribed from Table 4.

2/ See Table 1C for the estimates of $\overline{\sigma_{de}\delta_{de}\rho_{de}^{-1}}$.

-- Not Available

Table 3.--Sample Estimates of $\sigma_d^2, \sigma_{x(T)}^2$ and Ratios $\sigma_d^2/\sigma_{x(T)}^2$, $\sigma_X^2/\sigma_{x(T)}^2$, and σ_d^2/σ_X^2.

Housing Variables	Estimate of				
	σ_d^2	$\sigma_{x(T)}^2$	$\sigma_d^2/\sigma_{x(T)}^2$	$\sigma_X^2/\sigma_{x(T)}^2$	σ_d^2/σ_X^2
Owner occupied units	.0230	.2323	.0990	.9010	.1099
Renter occupied units.	.0230	.2323	.0990	.9010	.1099
Units with shared or no bath . . .	.0540	.1211	.4459	.5541	.8047
Units with bath for exclusive use	.0124	.1074	.1155	.8845	.1306
Sound.	.0783	.1530	.5117	.4883	1.0479
Deteriorating.	.0872	.1159	.7524	.2476	3.0388
Dilapidated.	.0298	.0520	.5731	.4269	1.3425

Source: Reference [17].

Table 4.--Sample Estimates of $\sigma_{de}, \sigma_{xy(T)}$, and Ratios, $\sigma_{de}/\sigma_{xy(T)}, \sigma_{XY}/\sigma_{xy(T)}$, and σ_{de}/σ_{XY}.

Housing variables (Dependent variable vs. independent variable)	Estimates of				
	σ_{de}	$\sigma_{xy(T)}$	σ_{de}/σ_{xyT}	σ_{XY}/σ_{xyT}	σ_{de}/σ_{XY}
Condition of Housing - Sound:					
vs. Owner occupied units	.0043	.01464	.2964	.7036	.4213
Renter occupied units.	.0014	-.03139	-.0446	1.0446	-.0427
Units with shared or no bath	-.0016	-.06600	.0244	.9756	.0250
Units with bath for exclusive use. . . .	.0149	.03573	.4167	.5833	.7144
Units with monthly rent less than $80. .	-.0003	-.02692	.0093	.9907	.0094
Units with monthly rent less than $60. .	.0004	-.02948	-.0136	1.0136	-.0134
Condition of Housing - Deteriorating:					
vs. Owner occupied units	-.0027	-.00497	.5493	.4507	1.2188
Renter occupied units.	.0014	.01662	.0818	.9182	.0891
Units with shared or no bath	.0010	.03275	.0302	.9698	.0311
Units with bath for exclusive use. . . .	-.0058	-.01973	.2955	.7045	.4194
Units with monthly rent less than $80. .	.0014	.01489	.0913	.9087	.1005
Units with monthly rent less than $60. .	.0007	.01538	.0481	.9519	.0505
Condition of Housing - Dilapidated:					
vs. Owner occupied units	.0007	-.00633	-.1106	1.1106	-.0996
Renter occupied units.	-.0010	.01476	-.0678	1.0678	-.0635
Units with shared or no bath	.0009	.03350	.0260	.9740	.0267
Units with bath for exclusive use. . . .	-.0045	-.01551	.2882	.7118	.4049
Units with monthly rent less than $80. .	.0004	.01253	.0295	.9705	.0304
Units with monthly rent less than $60. .	.0000	.01464	.0000	1.0000	.0000

Source: The sample of housing units used in the Content Evaluation Study of 1960 at the Bureau of the Census.

See reference [17] for further details.

Table 5.--Estimates of β, E_1 and β^*.

Housing variables (Dependent variable vs. independent variable)	Estimates of		
	β^* (1)	E_1 (2)	β (3)=(1)÷(2)
Condition of Housing - Sound:			
vs. Owner occupied units	.0630	1.2806	.0492
Renter occupied units	-.1351	.8625	-.1566
Units with shared or no bath	-.5454	.5680	-.9602
Units with bath for exclusive use	.3327	1.5164	.2194
Condition of Housing - Deteriorating:			
vs. Owner occupied units	-.0214	1.0981	-.0195
Renter occupied units	.0715	.9813	.0729
Units with shared or no bath	.2704	.5713	.4733
Units with bath for exclusive use	-.1837	1.2554	-.1463
Condition of Housing - Dilapidated:			
vs. Owner occupied units	-.0272	.8112	-.0335
Renter occupied units	.0635	.8438	.0753
Units with shared or no bath	.2766	.5689	.4862
Units with bath for exclusive use	-.1444	1.2426	-.1162

Source: Tables 3 and 4.

Table 6.--The Values of E_1 for a Set of Values of the Ratio σ_d^2/σ_X^2 and σ_{de}/σ_{XY}.

(The number of the upper level in the column labels indicate $\sigma_d^2/\sigma_{x(T)}^2$ and the ones on the lower level indicate σ_d^2/σ_X^2).

$\sigma_{de}/\sigma_{xy(T)}$	σ_{de}/σ_{XY}	$\sigma_d^2/\sigma_{x(T)}^2$ and σ_d^2/σ_X^2					
		.0000 .0000	.2000 .2500	.4000 .6667	.5000 1.0000	.6000 1.5000	.8000 4.0000
-.2000	-.1667	.8333	.666	.5000	.4167	.3333	.1667
.0000	.0000	1.0000	.8000	.6000	.5000	.4000	.2000
.2000	.2500	1.2500	1.0000	.7500	.6250	.5000	.2500
.4000	.6667	1.6667	1.3333	1.0000	.8333	.6667	.3333
.5000	1.0000	2.0000	1.6000	1.2000	1.0000	.8000	.4000
.6000	1.5000	2.5000	2.0000	1.5000	1.2500	1.0000	.5000
.8000	4.0000	5.0000	4.0000	3.0000	2.5000	2.0000	1.0000

Table 7.--Selected 1960 Census Data for Six Cities

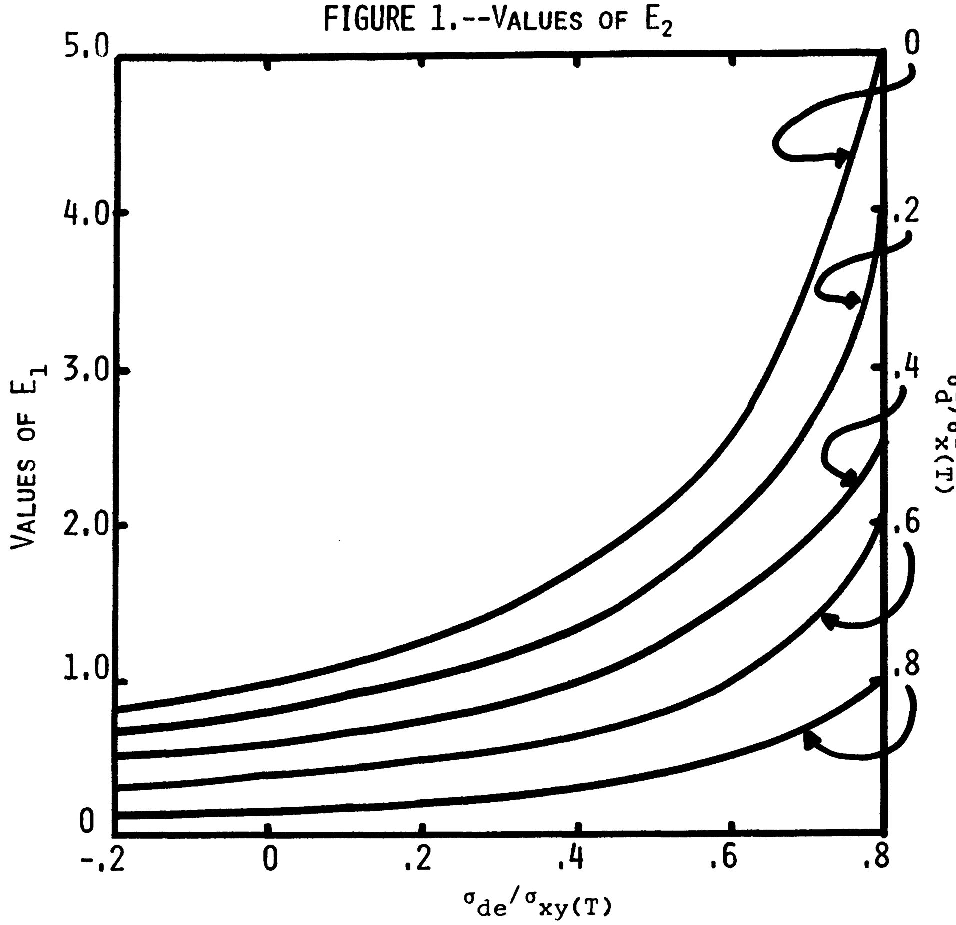
Subject	Camden, N.J.	Cleveland, Ohio	Ft. Wayne, Ind.	Louisville, Ky.	Shreveport, La.	South Bend, Ind.
1. Population	117,159	876,050	161,766	390,639	164,372	132,445
2. Housing units	37,015	282,893	53,002	128,238	54,191	42,590
3. Condition						
a. Percent deteriorating	15.3	14.1	11.8	14.7	16.4	10.5
b. Percent dilapidated	3.4	3.1	2.5	3.9	5.7	2.2
c. Percent substandard 1/	9.4	9.1	7.3	15.3	20.6	7.6
4. Area						
a. Census tracts 2/	27	203	39	111	40	34
b. Enumeration districts 2/	110	1,031	203	472	190	215
c. Blocks 2/	1,083	4,389	2,075	3,042	2,038	1,740

1/ Not an official Census classification. Used by other agencies. Includes "Sound" and "Deteriorating" units lacking one or more of these facilities; piped hot water, flush toilet for private use, bathtub or shower for private use, plus all "Dilapidated" units.

2/ Excludes areas in which there were no occupied housing units.

Source: p. 47 of Reference [18].

FIGURE 1.--VALUES OF E_2



REFERENCES

*The author gratefully acknowledges the useful comments given by Professor John Neter of the University of Minnesota, and Dr. Benjamin Tepping and Mrs. Barbara Bailer of the Bureau of the Census. Views expressed in this paper are those of the author and not necessarily those of the Bureau of the Census.

[1] Bailer, B.A., "The Effect of Varying the Size of the Interviewer's Assignment Area on ρ_d, the Interclass Correlation Between Response Deviations of Different Units Assigned to the Same Interviewer." Unpublished article, Center for Research in Measurement Methods, U.S. Bureau of the Census, September 28, 1967.

[2] Bailer, B.A., "Recent Research in Reinterview Procedures." Journal of the American Statistical Association, 63 March 1968.

[3] Berkson, Joseph, "Are there two regressions?", J. of Amer. Stat. Assoc., 45 (1950), pp. 164-180.

[4] Chai, John J., A Study of Effects of Correlated Measurement Errors and Grouping on the Ordinary Least Squares Estimator for the Regression Coefficient and Ordinary Estimator of the Product-Moment Correlation Coefficient of A Finite Bivariate Population. Ph.D. Thesis. University of Minnesota, 1968.

[5] Cochran, W.G., Sampling Techniques, Second Ed. New York: John Wiley and Sons, Inc., 1963.

[6] Durbin, J., "Errors in Variables," Rev. of Inter. Stat. Inst., 22 (1954), pp. 22-23.

[7] Goldberger, A.S., Econometric Theory, New York: John Wiley and Sons, Inc. 1964.

[8] Hansen, M.H., Hurwitz, W.N., and Madow, W.G., Sample Survey Methods and Theory, Vols. I and II. New York: John Wiley and Sons, Inc. (1953).

[9] Hansen, M.H., Hurwitz, W.N., and Pritzker, Leon, "The estimation and interpretation of gross differences and the simple response variance," Contributions to Statistics Presented to Professor P.S. Mahalanobis on the Occasion of His 70th Birthday. Calcutta, India: Pergamon Press(1964).

[10] Hansen, M.H., Hurwitz, W.N., and Bershad, M.A., "Measurement errors in censuses and surveys," Bulletin of the Inter. Stat. Inst., 38 (1961), pp. 359-374.

[11] Jessop, W.N., "One line or two?", Applied Stat., 2(1952),pp.131-317.

[12] Johnston, J., Econometric Methods. New York: McGraw-Hill Book Co., Inc. 1963.

[13] Mahalanobis, P.C., "Recent Experiments in Statistical Sampling in the Indian Statistical Institute", J. of the Royal Stat. Soc., 109(1946) pp. 18-55.

[14] Mandansky, A. "The fitting of straight lines when both variables are subject to error," J. Roy.Stat. Soc., 54(1959), pp. 173-205.

[15] Nair, K.R. and Banerjee, K.S., "A note on fitting of straight lines if both variables are subject to error," Sankhyā, Vol. 5(1943), p.331.

[16] Pritzker, Leon, "Response Response Research Branch Report No. 66", April 27, 1966, U.S. Bureau of the Census.

[17] U.S. Bureau of the Census. Evaluation and Research Program of the U.S. Censuses of Population and Housing, 1960: Accuracy of Data on Housing Characteristics, Series ER 60, No. 3. Dept. of Commerce, Bureau of the Census, Washington, D.C., (1964).

[18] U.S. Bureau of the Census, Measuring the Quality of Housing, An Appraisal of Census Statistics and Methods, Working Paper No. 25, Department of Commerce, Bureau of the Census, Washington, D.C. 1967.

STOCHASTIC MODELS OF CITY POPULATION DENSITY

Robert T. Amsden, Rutgers-The State University

Previous Studies

The literature which is relevant to the problem of the spatial distributions of urban populations may be classified in two ways. These serve either to describe directly urban population densities or to examine various stochastic processes which, although not necessarily directed toward the city, nevertheless, are relevant by providing components of a quantified mechanism. The descriptive articles bring out two major points. The first is that the various cities examined seem all to have a fairly consistent pattern of population density (where density implies members per unit area). This functional form may be approximated by a negative exponential where population density is primarily a function of radius from the center. Also, this work would seem to indicate that some cities eventually reach a point in time at which the inner or central population density reaches a plateau and then proceeds to fall off as the city grows and expands further. These few articles of description do not quantify any causes of the changes in the spatial distributions of urban populations.

No articles were found in the statistical literature which were directed toward the city problem and which developed city models incorporating random processes. However, a considerable amount of theory has been developed which deals with aspects of the problem. Much effort has been devoted to stochastic birth and death processes. Recently some attention has been given to stochastic processes which involve birth and death as well as the random position of the individual. These processes have been of two types: the first being random walk models in which an individual moves forward or backward a unit step (with random waiting times as an added complexity) while undergoing birth and death. Secondly there are diffusion type processes in which the individuals continually move infinitesimal distances backward or forward while subject to birth and death probabilities.

Bailey (1968) last year commented on the availability in the literature of various stochastic processes relevant to biological populations and then went on to say, "In general, the results obtained relate to the probability distribution of the total number of individuals present in any class. The possibility of a spatial arrangement of individuals is usually ignored or not explicitly introduced as it entails a considerable increase in complexity." In this article he does bring in the spatial aspect. Populations exist at points in one, two, or three dimensional space and the individuals are subject to random birth and death as well as random translation to the nearest neighboring populations. In discussing the need for more study about the joint probabilities of his models he comments that, "Most practical applications are of course more likely to involve two or three space dimensions, but it will probably be simpler to solve the one-dimensional problems first." In his conclusion he suggests that, "Methods are therefore required not only of characterizing the behaviour of the process for general discussion, but also for testing the model against observable data."

Studies of the Author

We begin with very simple, but illuminating, non-stochastic processes involving population movement. This idealized spatial distribution of population is acted upon by a purely deterministic force which moves individuals in an outward direction. It is shown that this kind of process, when acting upon a distribution of population, develops a complete hole in the center and suffers from lack of any fluctuation phenomena. Lest one be misled it is impossible to construct a model of city growth and spread based upon a purely deterministic force which will not develop a hole.

In order that we may allow some individuals to remain in the center of our model city and yet require others to leave, it becomes necessary to incorporate stochastic processes. A stratified model has been developed by the author in which we study the mean or expected number of individuals in each stratum as a function of time and initial conditions; individuals may move up (or outward in the city context) one stratum at a jump after a random waiting time. This model, which is built upon the analytical methods of Seal (1945) is a new development in that, in contrast to Seal's work, we start with initial numbers of individuals in each stratum and then proceed to examine how the averages of each stratum's count changes through time. We note that this structure lacks any fluctuation phenomena and that, although we have allowed random waiting times for individuals, once they are selected they can move only one step into the next higher (or outer) stratum.

A further step has been a new application of stochastic processes theory in which random waiting times and random distances of movement in the outward direction upon selection are applied to a sectionally continuous spatial distribution model of population. Feller and others have examined stochastic processes where the element is affected by random waiting times and random distances of movement upon selection; but there do not appear to be any publications which apply these concepts to an initially spatially distributed population of individuals and then examine the time path of the population density function. The author has so done and incorporated rates of birth and death and applied some of the analytical procedures outlined by Bartlett (1966, pp. 75-78). An expected population density function is derived in which we require differential elements of individuals to move after random waiting times and then by random translations once selected. This concept is not new, but the derivation has been carried through in its entirety so that we end up with a process in which the spatial population density function changes through time by means of the movement of differential elements.

Several applications, all of which are new, were then made of this expected population density process. Some of the explicit solutions involve modified Bessel functions. One case deserves special note. The spatial distribution of population on the plane is chosen to be initially in the uniform or rectangular form, as though individuals lived evenly distributed within a walled city with no one living outside. These individuals (or differential elements) are subjected to random waiting times, random movements in the outward direction upon selection, and rates of birth and death (at time zero the walls are flattened and the outward rush begins). An explicit solution is obtained, and it is shown that, except for time zero, the population density is always slightly higher slightly outside the center. The importance of such a result is that much, although not all, of the actual city data examined in the literature evidence central holes in population density. In the publications it has been suggested that this is a result of business interests concentrated in the center. The work of this author would suggest that such an effect might arise whenever there is a tendency of the inhabitants to move in an outward direction, especially if the city population density function is relatively flat near the center whenever this outward movement begins.

Up to this point we have considered individuals to be affected by random waiting times and random distances moved in the outward direction both of which are independent of the radial position of the subject. We then set up a model in which either of these random features is related to radial position. In either case we derive a Laplace transform which satisfies a differential difference equation. Explicit solutions are not derived.

Fluctuation phenomena are then brought into the models by examining the behaviour of a spatial distribution of discrete individuals. We begin by deriving the probability distribution for position of a single individual who is subjected to random waiting times as well as random distances moved in an outward direction once selected. The probability density for being at a point is then found as a function of time with the solution involving Bessel functions. A similar result is given by Feller (1966, p. 58). We build upon this process by next considering a spread out group of individuals each of whom is subjected to the above process. The expected density of individuals at a point is derived as a function of time. This simple derivation also represents a new development, and is important because the spatial aspect is brought into the model. However, it does not evidence any fluctuation.

In order to include fluctuation phenomena resulting from random waiting times and random translations, two new processes are developed in which we derive the probability distribution of counts of individuals inside a specified interval of position. A criterion is outlined for use of the model in city planning; a unit width of annular ring is defined for the city and then one asks how far from the center one can project the model before the coefficient of variation begins to blow up. Beyond this point the model would be inapplicable; but anywhere inside the ring, the city planner could use his model for projections of future population densities. In this work we have set up general expressions for the calculation of the coefficients of variation.

We should note several points. It is a fact that the movement of people through time in a city exhibits stochastic aspects. Introduction of stochastic phenomena into city models is the only way of reflecting the birth, death, and movements taking place in city development. Stochastic processes do account for major qualitative features described by the demographers. Finally we have established a criterion for determining the stability of our stochastic models.

REFERENCES

(1) Amsden, Robert T., Stochastic Theory of Population Density, (Dissertation), Rutgers-The State University, New Brunswick, New Jersey, 1969 (See this for a more exhaustive bibliography. Copies of the dissertation may be purchased through University Microfilms, Ann Arbor, Michigan).

(2) Bailey, Norman T. J., "Stochastic Birth, Death and Migration Processes for Spatially Distributed Populations," Biometrika (1968), 55, 189-198.

(3) Bartlett, H. S., An Introduction to Stochastic Processes with Special Reference to Methods and Applications, Cambridge University Press, London, 1966.

(4) Clark, Colin, "Urban Population Densities," Journal of the Royal Statistical Society, 114, P. IV (1951), 490-496.

(5) Feller, William, An Introduction to Probability Theory and Its Applications, II, John Wiley & Sons, Inc., New York, 1966.

(6) Seal, H. L., "The Mathematics of a Population Composed of Stationary Strata Each Recruited from the Stratum Below and Supported at the Lowest Level by a Uniform Annual Number of Entrants," Biometrika 33 (1945), 226-230.

ANALYZING PAIR DATA AND POINT DATA ON SOCIAL RELATIONSHIPS,
ATTITUDES AND BACKGROUND CHARACTERISTICS OF
COSTA RICAN CENSUS BUREAU EMPLOYEES

C. H. Proctor, North Carolina State University

Introduction

Linear statistical models have been widely used in sociology and even their more refined versions are being properly applied and wisely interpreted.[1] So perhaps it is a fitting time to create a little confusion with the expectation that the problems eventually will be ironed out. The source of the present difficulty is sociometric or interpersonal data. In particular, it is often found possible to characterize, with a variable value, the state of the relationship between two persons. The problem is how to investigate the association between such pair variables and the more customary point data on the same persons.

The issue has been quite a bit discussed under the title of "ecological correlations" for the case that groups of persons or households are analyzed as well as the persons or households. The notion of "nested models" in analysis of variance parlance with components of variance or intraclass correlation formulation can aid in interpreting such data but this viewpoint seems rather specialized. It views the aggregate as a sum of group components and individual components, and the variables measured on the individual parts are the same as those measured on the aggregate. In the setting we wish to work, the variables are different.

To make the problem hopefully easier to discuss we will take a concrete example. The data of this example were collected by self-administered questionnaires from the employees of the Costa Rican Census Office in 1952. There are three classes of variables that will be analyzed. These are:

(1) Three attitude-toward-work items,

(2) Interpersonal structure measured by two sets of sociometric choice data using the criteria: want-to-work-with and have-coffee-with, and

(3) Seven background variables (five sections of the office as represented by four indicator variables, schooling, age and sex).

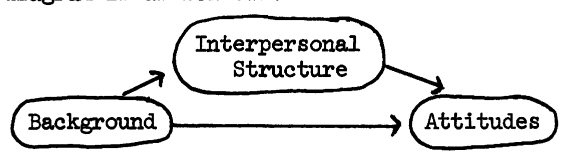
The interpersonal or sociometric variables were scored as mutual mention, one-way, or indifference for every pair of persons. From the 63 persons responding, there are 1,953 unordered pairs and 3,906 ordered pairs that can be formed. The other attitude and background variables were scored for each person and can be called "point" variables to contrast with the two sociometric or "pair" variables.

Preliminary Calculations

The initial approach to the data was pre-statistical, in that there was no thought of a probability mechanism, no parameters and thus no estimation. A regression analysis was done using the 1,953 pairs as the units of analysis. The three interpersonal structure pair variables were called: strength, direction, and unbalance and will be denoted z_1, z_2, and z_3, respectively. Their values were as follows: $z_1 = 0$ for indifference, $z_1 = 1$ for one-way and $z_1 = 2$ for mutual mention; $z_2 = +1$ when the first person mentions the first, $z_2 = -1$ for the reverse and $z_2 = 0$ if indifference or mutual; while $z_3 = 1$ if one-way and zero otherwise.

The point variables were used to form signed differences (first person's score minus second person's) and also to form absolute differences. These transformed variables then become pair variables derived from point variables. Thus there were 6 attitude plus 3 interpersonal plus 14 background equals 23 pair variables.

The causal mechanism that was posited had background variables giving rise to the attitude scores and also to the interpersonal structure with the structural variables further affecting the attitude scores. The path diagram is as follows:



Thus, two sets of regression equations for the pair variables were examined:

(1) Both attitude and structure on background;

(2) Attitude on structure with both adjusted for background.

When the coefficients were examined and when interpretations were attempted the results made good sense but there was also a good bit of nonsense introduced by the naive approach. Several points emerged:

(a) The signed differences and direction (z_2) variables are both asymmetric, in that they change sign when the order of the pair is reversed, but the absolute differences along with strength (z_1) and unbalance (z_3) are symmetric. That is, they do not depend on the order of the pair. These two kinds of variables ought not be mixed in the same regression analysis.

(b) Although it was interesting to see how background affects structure, it does not seem advisable to "adjust" structure for background when examining how structure affects attitudes. This is a substantive question of outlook to the causal mechanism.

(c) The F-ratios and regression coefficient t-test statistics under the independent observation model between the signed differences (attitudes on background) were stupendously large and do require to be corrected.[2] The regression coefficients themselves are, however, numerically equal to those of the point variable analysis.

(d) The regression of structure on background may be calculated and interpreted under the usual model, namely fixed independent variable with independent homogeneous error, but counting 1,953 observations.

(e) Using absolute differences makes the model equation assumptions problematic and this will be treated below.

The computations were then redone by first regressing the attitude point variates on the background point variables and then calculating residual attitude point scores. These residuals were used to compute absolute difference pair data which were then regressed on strength and direction, while the signed difference of residuals pair data was regressed on direction. The regression coefficients of these three analyses appear in Table 1.

The regression coefficients, and in this instance even the accompanying t-values, that were produced by the computer can be used to screen relationships. Under this criterion attitude differences over whether coworkers should be friends showed dependence on both the coffee and the work interpersonal structures. When the strength of the relationship increased, the attitude differences decreased. The distribution of absolute differences themselves are shown by coffee relationship strength class in Figure 1. Also shown there are the log (Absolute difference + .1) transforms. This log transform was suggested by an argument to be presented shortly and appears to have stabilized the within-class distributions.

The analysis of variance of the log transformed absolute differences shows that $\bar{y}_0 = -.81$, $s_0^2 = .53$ for $n_0 = 1917$ pairs at zero strength; $\bar{y}_1 = -.89$, $s_1^2 = 78$ for $n_1 = 19$ pairs (one-way coffee choices); and $\bar{y}_2 = -1.50$, $s_2^2 = .62$ for $n_2 = 17$ pairs of strength 2 (mutual coffee choices). The pooled estimate of error variance was $s_p^2 = .5316$. Thus the differences themselves gradually decreased with an increase in the strength of the relationship. The distributions are reasonably well behaved as Figure 1 shows and the variance homogeneity is encouraging. Incidentially, the analysis of variance for the absolute differences led to $F = 3.91$ while for the log transformed differences $F = 7.69$ for the same case of three levels of strength. This difference in F values could be due to the large population fourth moment of the untransformed data, and an unexpectedly large overestimate of the error variance.

Statistical analysis

To interpret these means and variances some stochastic model needs to be suggested. There seemed to be two main approaches to explaining the data. Either census bureau employees begin by taking a variety of stands on the attitude question and then move together if their relationship is close or they begin by being initially identical and move apart if their relationship is not close. The model we will use supposes a process of random separation or drifting apart of two people that is modified by a gradual passage to uniformity or agreement if a relationship exists.

Although the following derivation is historically faithful to its conception and is given to bolster its use, the model should not be judged exclusively by this line of reasoning. Let $d_{ij}(t)$ be the difference in attitude scores between two persons i and j at time t. This has a sign. $d_{ij}^2(t)$ is a more appropriate measure of distance as it has no sign, and will be used in the following. When dealing with quantities that have no sign the subscripts i,j will be taken with $i < j$ (only the upper triangle of the matrix is used). Now $d_{ij}^2(t+1)$ should be related to $d_{ij}^2(t)$ in some way. If this relationship is represented multiplicatively as:

$$d_{ij}^2(t+1) = \beta d_{ij}^2(t)\epsilon_{ij}(t+1)$$

$$= \left(e^{2\alpha X_{ij}}\right) d_{ij}^2(t) \left(e^{2\delta_{ij}(t+1)}\right)$$

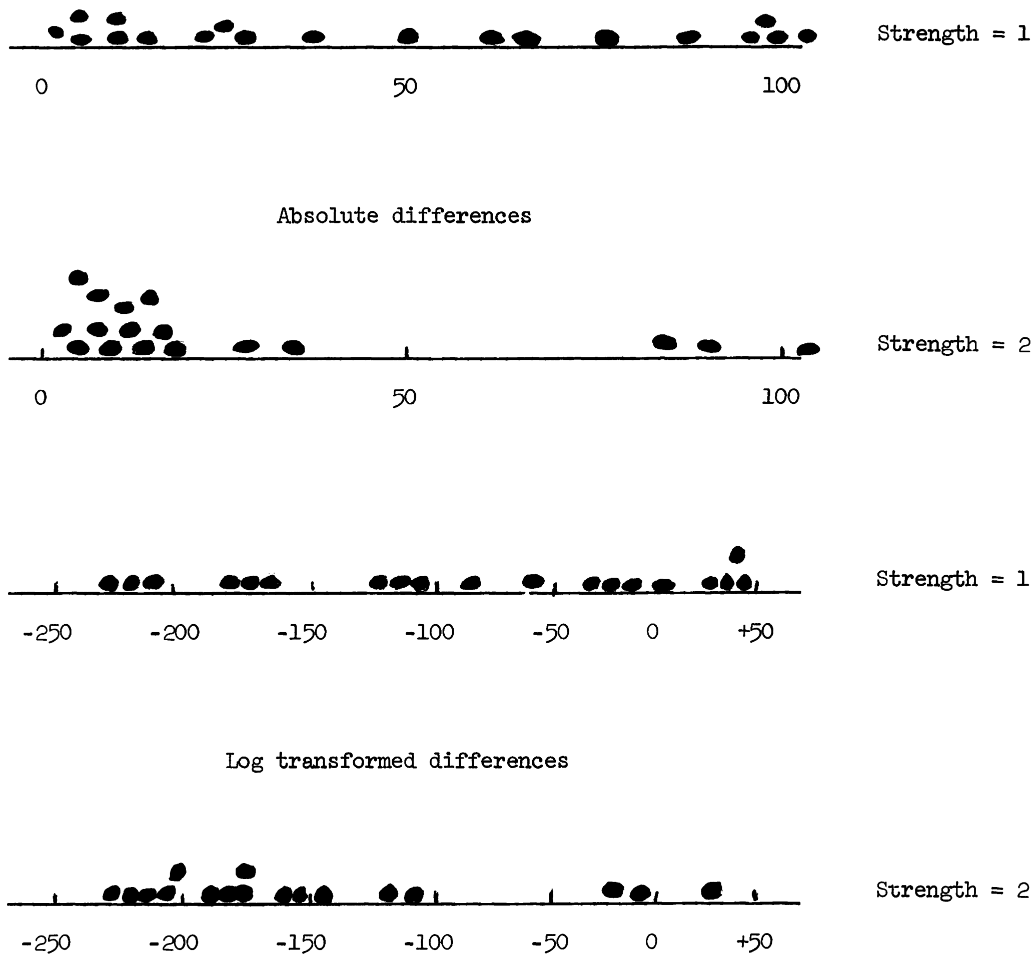
where X_{ij} is the strength of the relationship between i and j and $e^{2\alpha X_{ij}}$ is the systematic multiplier while $e^{2\delta_{ij}(t+1)}$ is the random

Table 1. Regression coefficients for pair analyses, boxed coefficients exceed 2 standard errors.

Regression Run	Independent Variable	Dependent Variables are indices of extent of work-rational orientation to: Salary Allocation	Explaining Mistakes	Friendship among Co-workers
I. Point Regression:	Jefes	-.076	-.120	.451
	Secretaries	.328	.117	.219
	IBM	.323	-.339	.170
	Coding	.251	.379	-.052
	Age	-.071	.106	.007
	Sex	-.109	-.023	-.195
	School	.038	-.079	-.033
II. Pair Regression:				
Work Pairs	Strength	-.043	.152*	[-.105]
	Unbalance	.069	-.199	-.015
Coffee Pairs	Strength	-.073	.046	[-.098]
	Unbalance	-.078	-.164	.131
III. Pair Regression:				
Work Pairs	Direction	.034	.010	-.118
Coffee Pairs	Direction	-.137	.101	.113
Separate Pair Regressions:				
Work	Strength	-.060	.164*	[-.129]
Coffee	Strength	-.079	.070	[-.121]

*These coefficients were close to twice their standard errors.

Fig. 1. Distributions of absolute differences and of log transformed differences by strength of coffee pair relationship.



one. Taking logs and cancelling 2's leads to:

$$\log| d_{ij}(t + 1) | = \alpha X_{ij} + \log | d_{ij}(t) | + \delta_{ij}(t+1),$$

or by rewriting the terms in an obvious way ($\log| d_{ij} | = D_{ij}$):

$$D_{ij}(t + 1) = \alpha X_{ij} + D_{ij}(t) + \delta_{ij}(t + 1).$$

In order to further normalize the distribution of the δ_{ij}'s the values $y_{ij} = \log(|d_{ij}| + .1)$ were analyzed but the discussion will proceed with D_{ij}. Moving toward a slight bit more of generality we can dispense with α as the single linear regression coefficient and use an effects model as:

$$D_{ij}(t + 1) = \alpha(X_{ij}) + D_{ij}(t) + \delta_{ij}(t + 1)$$

in which α is a function; namely:

$$\alpha(0) = 0$$
$$\alpha(1) = \alpha_1$$
$$\alpha(2) = \alpha_2.$$

If it be supposed that $V(\delta_{ij}(t)) = \sigma^2_{ij}$ for all t and the δ_{ij}'s are independent from one time to the next then we can follow the course of $D_{ij}(t)$ through time as a function of X_{ij} and deduce some convenient distributional properties. If $X_{ij} = 0$ and we take to begin the process $D_{ij}(0) = D_0$ say, then

$$D_{ij}(1) = 0 + D_{ij}(0) + \delta_{ij}(1) = D_0 + \delta_{ij}(1)$$
$$D_{ij}(2) = 0 + D_{ij}(1) + \delta_{ij}(2) = D_0 + \delta_{ij}(1) + \delta_{ij}(2)$$
$$\vdots$$
$$D_{ij}(t) = D_0 + \sum_{k-1}^{t} \delta_{ij}(k) .$$

However if $X_{ij} = 1$ or 2 then (taking $X_{ij} = 1$ for example)

$$D_{ij}(1) = \alpha_1 + D_0 + \delta_{ij}(1)$$
$$D_{ij}(2) = \alpha_1 + \alpha_1 + D_0 + \delta_{ij}(1) + \delta_{ij}(2)$$
$$\vdots$$
$$D_{ij}(t) = t\alpha_1 + d_0 + \sum_{k=1}^{t} \delta_{ij}(k) .$$

Similarily:

$$D_{ij}(t) = t\alpha_2 + D_0 + \sum_{k=1}^{t} \delta_{ij}(k), \text{ for } X_{ij} = 2.$$

For data that are from a one-point-in-time survey, such as are those of the present example, one cannot know t. It could be supposed that attitude reassessment occurs very frequently so that t (the number of reassessments) is very large and then we would suppose that α_1 and α_2 are very small but as t increases $t\alpha_1 \rightarrow \lambda_2$ and the variances of each of the $\delta_{ij}(t)$ also are very small so that again as t increases $V(\sum_{k=1}^{t} \delta_{ij}(k)) \rightarrow \sigma^2_\delta$.

the model equations then become:

$$D_{ij}(\text{survey time}) = D_0 + \delta_{ij}, \quad \text{if } X_i = 0$$

(1) $$D_{ij}(\text{survey time}) = D_0 + \lambda_1 + \delta_{ij}, \quad \text{if } X_{ij} = 1$$

$$D_{ij}(\text{survey time}) = D_0 + \lambda_2 + \delta_{ij}, \quad \text{if } X_{ij} = 2$$

with $V(\delta_{ij}) = \sigma^2_\delta$.

These equations have the flavor of an ANOVA model. Before using the F-tests suggested by this formulation, a problem would seem still to be the dependence among the δ_{ij}. Here the patterns are numerous but still finite. The major departure could be expected as a correlation between a certain δ_{ij} and the associated ones of the forms $\delta_{ij'}$ or $\delta_{i'j}$ where the primed subscript is unequal to the corresponding unprimed one.

Since there are 1953 pairs there become 1,906,128 pairs of pairs. Of these, 119,133 or about 6.3% have a subscript in common. Table 2 shows twelve types of pairs of pairs (the upper number is of mutually exclusive pairs and the lower gives the overlapped pairs) as counted from the coffee structure.

If such seems reasonable, it may be supposed that the δ_{ij}'s of overlapped pairs of pairs were correlated while the mutually exclusive pairs were not. To investigate such a possibility, a sample of 10 pairs from Table 2 was drawn at random from pairs with strengths 1 and 2 (See Table 3). The 45 pairs of pairs were found to include 39 mutually exclusive and 6 overlapped. The ten residuals (y_{ij} minus fitted y_{ij}) were recorded and from them the differences were computed and then the variances of differences of the two types were calculated separately. These data are given in Table 4, along with the original data for strengths 1 and 2 in Table 3. This gave:

Table 2. Counts of Non-overlapping and Overlapping Pairs of Pairs by Combination of Strengths of the Pair (number before plus sign is of non-overlapping pairs of pairs)

Strength of One Pair	Strength of Other Pair			No. of Pairs
	0	1	2	
0	1,723,122 +113,364			$n_0 = 1917$
1	32,788 +3,635	163 +8		$n_1 = 19$
2	30,428 +2,161	306 +17	118 +18	$n_2 = 17$

$s^2_{ex.} = .8373$, and $s^2_{lap.} = .0815$, respectively. If δ_{ij} and $\delta_{i'j'}$ are mutually exclusive it is supposed that $V(\delta_{ij} - \delta_{i'j'}) = 2\sigma^2_\delta$ but if not, then $V(\delta_{ij} - \delta_{i'j'}) = V(\delta_{ij} - \delta_{ij'}) = 2\sigma^2_\delta(1-\rho)$. One could, therefore use $.0815/.8373 = .0973$ as an estimate of $1 - \rho$. Thus $\hat{\rho} = .90$, quite a sizeable correlation and worthy of special attention.

To obtain unbiased estimates of the parameters D_0, λ_1 and λ_2 is straight forward using the mean values from the three classes (recalling $\bar{y}_0 = -.81$, $\bar{y}_2 = -.89$ and $\bar{y}_2 = -1.50$). The results were: $\hat{D}_0 = -.81$, $\hat{\lambda}_1 = -.89 + .81 = -.08$, $\hat{\lambda}_2 = -1.50 + .81 = -.69$.

The crucial parameter in representing the effect of social structure on attitudes would seem to be λ_2 whose estimate is $\hat{\lambda}_2 = -.69$. To test if $\lambda_2 = 0$ is tantamount to deciding if there is an effect. The suggested test procedure is as follows.

It can be fairly judged that $\hat{\lambda}_2$ will be covered by the central limit theorem. To estimate its variance one must take into account the covariances of overlapped pairs of pairs. These arise in three places in each of $V(\bar{y}_0)$ and $V(\bar{y}_2)$ as well as in $\mathrm{Cov}\,(\bar{y}_0, \bar{y}_2)$. The following computation is based on Table 2.

$$V(\bar{y}_0 - \bar{y}_2) = V(\bar{y}_0) + V(\bar{y}_2) - 2\,\mathrm{Cov}(\bar{y}_0, \bar{y}_2)$$

$$= \left[\left(\frac{1}{1917} + \frac{2(.9)113,364}{1917^2}\right) + \left(\frac{1}{17} + \frac{2(.9)18}{17^2}\right) - \frac{2(.9)2161}{1917 \cdot 17}\right][.532]$$

$$= \left[\left(\frac{1}{1917} + \frac{1}{17}\right) + 2(.9)\left(\frac{113,364}{1917^2} + \frac{18}{17^2} - \frac{2161}{1917 \cdot 17}\right)\right][.532]$$

$$= [.10762344][.532] = .05725567$$

As an example, consider the expression for $V(\bar{y}_0)$. The estimate $\bar{y}_0$ itself is the mean of 1917 quantities, but not independent quantities. From Table 2 one finds that 113,364 pairs are correlated.

Table 3. Values of $y_{ij} = \log(|d_{ij}| + .1)$ for Strengths 1 and 2.

Strength 1 (i,j)	y_{ij}	Strength 2 (j,j)	y_{ij}
3,6	-2.18	4,27	-2.18
5,9	-2.01	7,16	-1.32*
5,54	- .01	10,40	-2.24
8,34	+ .09*	10,50	-2.18*
10,27	-2.07	11,13	-1.69*
15,26	- .94	11,14	-2.02*
16,43	-1.71	11,54	-1.49*
18,25	+ .10*	13,14	-1.88*
20,41	-1.75	13,54	-1.97
23,56	-1.15*	14,54	-1.64
28,43	-1.30	15,18	+ .19
29,61	- .19	15,56	-1.87
30,55	- .37	30,42	-1.83
35,41	- .53	30,58	- .11
37,43	+ .13	32,51	-1.12
42,55	- .29	40,50	-2.12*
46,52	-1.79	42,58	- .05
47,62	+ .20		
55,58	-1.19		

$n_1 = 19$ $\qquad$ $n_2 = 17$

$\bar{y}_1 = -.8096$ $\qquad$ $\bar{y}_2 = -.8926$

*Pair sampled to estimate ρ.

Table 4. Differences (in absolute values) of residuals from Table 3 for overlapping pairs of pairs and for mutually exclusive pairs of pairs.

Mutually Exclusive					Overlapping
.01	1.60	1.61	.86	.69	.06
1.24	1.25	.44	.37	.30	.33
.80	.81	.42	.70	.43	.20
1.66	1.67	.07	.17	.10	.19
1.17	1.18	.26	.56	.39	.53
1.50	1.51	.27	.80	.63	.14
.97	.98	.12	.49	.24	
1.36	1.37	.36	.16		

Sum of Sqs. = 32.6538 SS = .4891

Average squared difference = .8373 Av. Sq. Diff. = .08152

****

Thus

$$V(\bar{y}_0) = V\left[\frac{1}{1917}(y_1 + y_2 + \cdots + y_{1917})\right]$$

$$= \left(\frac{1}{1917}\right)^2\left[\sum_i V(y_i) + 2\sum\sum_{i<j} \mathrm{Cov}(y_i,y_j)\right]$$

$$= \left(\frac{1}{1917}\right)^2 [1917\,\sigma_\delta^2 + 2(113,364)\rho\,\sigma_\delta^2]$$

The standard error of $\hat{\lambda}_2$ then is S.E. $(\hat{\lambda}_2)$ = .2393 and a test of $\lambda_1 = 0$ may be based on the critical ratio of .69/.2393 = 2.88. There is evidence that $\lambda_1 \neq 0$ from the fact that the probability of a standard normal deviate exceeding 2.88 is about .002.

The previous analysis (suggested estimates and tests) appeals to an analogy with the analysis of variance. The model equations are the same except for the dependence among the δ_{ij}'s. When that dependence is characterized by a non-zero covariance ($=\rho\sigma_\delta^2$ say) between δ_{ij} and δ_{kh} when the sets $\{i,j\}$ and $\{k,h\}$ overlap, but zero otherwise then some interesting problems arise concerning the covariance matrix of the δ_{ij}'s. Let the single (generic subscripts r and s replace the pair subscripts (i,j) and (k,h) in which the range of r and s is 1, 2, ..., $n(n-1)/2(= m$ say), and let the ordering on r and s be such that $r < s$ if $i < k$ or when $i = k$ if $j < h$ (recall that $i < j$ and $k < h$ by convention). Now the m by m covariance matrix of the $\delta_{ij} = \delta_r$'s will be denoted σ_δ^2 V where V contains 1's along the main diagonal and a scattering of ρ's and 0's off the main diagonal.

The model equations (1) can be rewritten in matrix notation as:

(2) $\underset{\sim}{y} = X\underset{\sim}{\theta} + \underset{\sim}{d}$ with $E(\underset{\sim}{d}) = \underset{\sim}{0}$ and $E(\underset{\sim}{d}\,\underset{\sim}{d}') = \sigma_d^2 V$.

Here y and d are m x 1 and X is m x p where p is the number of parameters.

In the particular example given above $\underset{\sim}{\theta}' = (D_0, \lambda_1, \lambda_2)$ and the 1,953 by 3 incidence matrix X consisted of 0's and 1's. The estimate of $\underset{\sim}{\theta}$ given by least squares theory is :

$$\hat{\theta} = (X'V^{-1}X)^{-1}X'\,V^{-1}\,Y$$

The naive estimators suggested above are these with V = I.

For the present data such estimates (with $V \neq I$) were not calculated, and this for two reasons. The size of V is such as to make its inversion a doubtful computing operation and also the value of ρ was not known. Consequently, it is not known how closely the naive estimates $\underset{\sim}{\hat{\theta}} = (-.81, -.08, -.69)$ are to the least squares ones or, more to the point, how the variances of the naive estimates compare to those of the least squares ones.

Although the naive procedures may soon become replaced by the least squares or even better ones, let us try to cast them in a slightly more general notation. The naive estimates are $\bar{\theta} = (X'X)^{-1} X'y$ as mentioned above. The p by p covariance matrix of $\bar{\theta}$ is given by:

$$E(\bar{\theta}\,\bar{\theta}') = \sigma_d^2 (X'X)^{-1} X' \; V\, X(X'X)^{-1} = \Sigma_{\bar{\theta}} \text{ say.}$$

For the usual ANOVA arrangements of X the entries of this matrix can be calculated using counts of overlapped pairs as in Table 1 above. The efficiency of the method of sub-sampling the pairs as described above and the calculation form for estimating ρ is not known. It was dictated by my limited computing resources. For example, the estimate of ρ cannot be expected to be unbiased even though the estimates of σ_δ^2 and $(1 - \rho)\, \sigma_\delta^2$ are. The estimates for these based on $s_{ex.}^2$ and $s_{lap.}^2$ are themselves not unbiased due to some slight negative covariances among the residuals.[4]

FOOTNOTES

1. This comment was strongly prompted by seeing the paper by W.H. Sewell and V.P. Shah, "Parent's education and children's educational aspirations and achievement," Am. Soc. Rev., 33:191-209, April 1968.

2. The factor involved appears to be the multiplier $(n\text{-}p\text{-}1)/[\binom{n}{2} - p - 1]$ which is to be applied to the F-ratios and its square root to the t-values of the paired variables analysis to obtain those of the single variable analysis. Here p is the number of independent variables in the regression.

3. Kendall, M.G. and Stuart, Alan. The Advanced Theory of Statistics, Vol. 2, New York, Hafner, 1961, p. 87.

4. Anscombe, F.J., "Examination of Residuals," in Proceedings of the Fourth Berkeley Symposium, University of California Press, 1961.

ATTITUDE FORMATION AS A FUNCTION OF INFORMATION THROUGH TIME*

Arnold Gordon, Northeastern Illinois State College

We are concerned with the history of a specific attitude (such as hostility) as it determines an individual's social interaction. The social response of one individual to another, or to a group, may depend on the readings which the individual has from various indicators reflecting the attitudes of others. Social response may also be a function of how these readings are retained.

This paper will attempt mathematical models (suggested by Milton Friedman's economic model for expected income)*** to describe the genesis of an emotional attitude, through time, as the subject is exposed to varying amounts of information. The models will be deterministic and stochastic. ****

May one predict, for example, what part (or proportion) of felt hostility will be expressed by an individual in units of time? In a day, week, or month, the environment moves attitudes in each of us. Hostility adds in units of time. Also, engendered hostility is reduced (forgotten, decays) in symbolic and overt individual responses in these same units of time.

This submission is in two parts: Model One suggests that current measurement of an attitude is most influenced by the bits of information historically closest to the measure.

Model Two suggests formulas for following and predicting attitude development.

Model One

Let us assume what appears logical: an attitude is most influenced by its most recent experiences.

Let H be hostility experiences in any time interval. H in one interval is equal to H in any other time interval.

We call the present time point of measurement T. The unit of time, of which T is the end point, is called t. t-1 precedes t.

Let i be a t period influenced by H. $i = 1, 2, 3 \ldots. n$, such that, $i \leq n$.

Let the proportion of H retained or forgotten in the i th interval, after the t interval in which it occurred, be Pi.

There are numerous possible t period combinations and their associated Pi proportions (or ratios). The mean, M, of these possible combinations is:

$$M = \sum_{i=1}^{n} P_i \, i \qquad (a)$$

Let the values of H, measured at T, be V_T, so that:

$$V_T = H_t(1) + H_{t-1}(1 - P_1) + H_{t-2}(1 - P_1 - P_2) + \ldots + H_{t-n+1}(Pn) \qquad (b)$$

V_T is the historical influence of Pi proportions up to the present moment, T. We note in formula (b) (if H_t (1) represents 100% of hostility feeling at time T) that each succeeding Pi unit subtracts larger and larger amounts from $H_t(1)$, therefore, remoter time units carry less and less weight.

Note that (a) and (b) are deterministic formulas. The H influence is not random and H is constant over time with zero variance between periods. Pi represents constant proportions through time so that, if P_1 is 50%, then 50% of H decays in the period immediately before the interval in which H_t was made.

Psychology, with its usual nominal and ordinal measuring scales, would appear unfit territory for deterministic approaches. Later we shall deal with a stochastic theory.

Since, by assumption, $H_t = H_{t-1} = H_{t-n}$, from (b) we find:

$$V_T = H \left[1 + (1 - P_1) + (1 - P_1 - P_2) + \ldots + Pn\right] \qquad (c)$$

and

$$V_T = H \left[(P_1 + P_2 + P_3 + \ldots + Pn) + (P_2 + P_3 + \ldots + Pn) + (P_{n-1} + Pn) + Pn\right] \qquad (d)$$

Pn is in n expressions and Pn-1 is in n-1 expressions, therefore:

$$V_T = H \left[1 P_1 + 2P_2 + 3P_3 + \ldots. + n\,Pn\right] \qquad (e)$$

and from (a):

$$V_T = H \left(\sum_{i=1}^{n} P_i \, i \right) = HM \qquad (f)$$

The memory schedule for hostility, H, refers to the proportions of H in the various categories. The proportions of H in one t period over V_T is:

$$H_t (1) / HM \qquad (g)$$

$$H_{t-1} (1 - P_1) / HM \qquad (h)$$

$$H_{t-2} (1 - P_1 - P_2) / HM \qquad (i)$$

. . . . . .

$H_{t-n} + 1 \,(Pn) \,/\, HM$ (j)

Since H is constant, we may cancell and add:

$(P_1 + P_2 + P_3 + \ldots + Pn) \,/\, M$ (k)

$(P_2 + P_3 + \ldots + Pn) \,/\, M$ (l)

$(P_3 + \ldots + Pn) \,/\, M$ (m)

$(Pn) \,/\, M$ (n)

In the addition of each numerator of formulas (k) through (n) we have the proportion of remembered hostility from various time periods, which is what remains after forgetting. Apparently, we may achieve as many different estimates of hostility summations as there are Pi retention terms. Recall that each i category in the retention schedule (between the parenthesis signs, above) represents Pi's which are first responding to the influences of an earlier H time interval. Pi is that proportion of total hostility which is still felt, or forgotten, in one unit of time.

For example:

1. Assume each new, different, hostility experience to be totally felt (equals 100%).

2. Assume each t time period to possess its own new, 100% hostility (H).

3. Assume H is forgotten at the following rate:

   $t - 1 = 50\% = P_1$

   $t - 2 = 30\% = P_2$

   $t - 3 = 20\% = P_3$ then,

t	H	t-1	t-2	t-3	Sum of t-1 t-2 t-3	V_T
1	100					100
2	100	50			50	150
3	100	50	30		80	170
4	100	50	30	20	100	170
5	100	50	30	20	100	170
6	100	50	30	20	100	170

Starting at level t_3, we note the vertical V_T column is the same number, 170, as those below it. We suggest that the normal human animal eventually reaches some level of healthy hostility control. A level where incoming conscious hostile experiences balances those forgotten.

$P_1 + P_2 + P_3$ (in t_3 above) estimates the distribution of hostility for single time periods:

$(0.5)(1) + (0.3)(2) + (0.2)(3) = 1.7$

The illustration assumes new hostility to every i period. Skipped, or infrequent H, suggests that residue hostility may completely decay.

Evidently the younger memory categories (t-1 is the most recent; t - n + 1, the oldest are weighthier through time relative to older memories. This should indicate that our immediate hostile experiences are more potent to retention than older ones.

In all the above work we have assumed determinism, stability, and constancy of i, H, t, and Pi. Psychology, however, is a science of random variables. What is control and prediction in it is correlation and values varying within and v 'thout significance levels. Once we accumulate sufficient information on past V_T values and assume normality, we may then set variance limits about the mean.

We assume, briefly, one stochastic approach as follows: Let H be a random variable in one t period of varying size. Suppose we have only two possibilities for i, one and two, i.e., hostility decays after one period or after two.

Assume, also, that the proportion decaying after one can be either 0.4 or 0.6 with a probability of one-half for each. Hostility which does not decay after one must, in this pretense, decay after two. Also, for 2, there is a 0.5 probability that 60% will be forgotten after two periods and 0.5 that 40% will decay after two periods.

Given the above, the mean hostility period equals:

$$0.5 \left[1\,(0.4) + 2\,(0.6)\right] + 0.5 \left[1\,(0.6) + 2\,(0.4)\right] = \quad (o)$$

$$1 \left[0.5\,(0.4) + 0.5\,(0.6)\right] + 2 \left[0.5\,(0.6) + 0.5\,(0.4)\right] = \quad (p)$$

$$1\,(0.5) + 2\,(0.5) = 1.5 \qquad (q)$$

(o) computes the mean, using the previous assumptions, as follows: There are two probability distributions, each having one-half weight (probability) in the total. In the first distribution, 40% of felt hostility is forgotten after one period, 60% after two periods. For the second distribution, 60% decays after one period, 40% after two. The two combined distributions terminate in (q); (q) demonstrates 50% retention after one period and 50% retention after two.

Sigmund Freud's approach (early childhood experiences are vital to personality development), and the current Harvard University studies on the first three years of life (Drs. Burton White and Jerome Bruner), appears against our hypothesis that the most recent attitude experiences are the most influential. Future investigations might pursue this: techniques or mechanisms for handling attitudes are formed as a function of our earliest experiences. When and how these mechanisms are called in to use is a function of our most recent experiences.

Model Two

Let us assume a person has a psychiatrically approved(normal, healthy, adjusted) optimal way of handling hostile feelings. He may then behave non-optimally in two extreme forms: rigid and repressed on one hand, violently anti-social on the other. It is an obvious convenience for psychologists to be able to trace and predict future deviant behavior. Numerous instruments describe personality at the time of measurement. Model Two proposes an ongoing series of correlated personality tests, administered at stated intervals, sufficient to indicate a trend. It may also proclaim, to therapist and counselor, which single time unit is noteworthy in abnormal behavior.

The following is untested theory:

1. The ratio of overtly expressed hostility provides a personality measure of hostility equilibrium, or control, in a unit of time t.

   $$m = H_1 / H_2 \qquad (r)$$

   where, m = proper handling of hostility in one unit
   H_1 = overt expression
   H_2 = memory hostility

2. The ratio of expressed hostility (H_1), to present hostile external stimuli (H_{22}), is also a measure of personality control in one unit of time t.

   $$m = H_1 / H_{22}$$

3. Therefore, $m = H_1 / H_2 + H_{22}$ (s)

4. The aging schedule, M, or what happens to hostility through time, may be an effective way of currently handling and predicting personality deviations.

5. The mean average of $H_2 + H_{22}$, in many time periods, is a predicting measure of H_1 in one period for normal individuals.

6. Let R_2 equal $H_2 + H_{22}$ in any time interval. Let R_2 in one interval be equal to R_2 in any other interval.

7. One time period may be a day, week, month, etc.

8. Let m (number 3, above) in the i th. interval be Pi. Pn indicates m through n periods. There are, theoretically, unlimited combinations of Pi patterns starting and interacting in different t periods.

9. The mean, M, of all Pi periods is:

   $$M = \sum_{i=1}^{n} P_i\, i \qquad (t)$$

10. Finally, let the total value of H_1 at time T be V_T.

The above purports to be a deterministic model where the exogenous variables are not random. R_2 is constant and equal in different t's with variance zero. Pi is a constant proportion in different intervals, not a probability. In equilibrium, R_2 for t equals R_2 for t + 1, t + 2, etc.; also, Pi at t equals Pi at t + 1, etc.

$$V_T = R_{2t}\,(P_1) + R_{2t+1}(P_2) + R_{2t+2}\,(P_3) + \ldots + R_{2t-n+1}\,(Pn) \qquad (u)$$

The first term to the right of the equal sign signifies R_2 for period t, part or all of which may transform into H_1. The second term includes residue from P_1 in P_2, plus new R_2 for period t + 1. The third term includes residue from $P_1 + P_2$ plus new R_2 for period t + 2, and so on.

Since, by assumption, $R_{2t} = R_2 t + 1$, expression (u) is rewritten as:

$$V_T = R_2 \left[(P_1 + P_2 + P_3 + \dots + Pn - 1 + Pn)\right] \quad (v)$$

Because Pn is equal in n expressions, V_T is rewritten:

$$V_T = R_2 (1P_1 + 2P_2 + 3P_3 + \dots + nPn) \quad (w)$$

Substituting formula (u) in (x):

$$V_T = R_2 \left(\sum_{i=1}^{n} P_i i\right) = R_2 M \quad (x)$$

and

$$M = V_T / R_2 \quad (y)$$

and

$$R_2 = V_T / M \quad (z)$$

Equation (x) proposes that hostile expression, H_1, is a function of past retention.

Equation (y) suggests that, in equilibrium, an effective control of hostility depends upon the reduction of R_2 tension through time.

Formula (z) indicates that average, over-all hostility, may demonstrate the effectiveness of adjustment in a single time unit.

We suggest the Minnesota Multiphasic Personality Inventory (adapted) as an instrument for Model Two measurement in single time periods: One of the secondary scales, hostility (Ho).

---

* Presented at the August, 1969, annual meeting of the American Statistical Association, New York City

** Psychology Department, Northeastern Illinois State College, Chicago, Illinois

*** Milton Friedman, A Theory of the Consumption Function, Princeton University Press, Princeton, 1957, Chapter 3.

**** The writer is grateful for the critical evaluations of Haskel Benishay, Professor of Managerial Economics, Northwestern University. The writer alone is responsible for judgement errors.

THE 1970 SOVIET CENSUS OF POPULATION: CONTENT, ORGANIZATION, AND PROCESSING

Frederick A. Leedy, U.S. Bureau of the Census

On January 15, 1970, the Soviet Union will begin taking its fifth census of population. Soviet law has no requirements for regular censuses, and the four previous enumerations of the population in the U.S.S.R. were taken at varying intervals--1920, 1926, 1939, and 1959. The one comprehensive census of population in Russia prior to the Soviet period was taken by the Tsarist Government in 1897.

Plans for the 1970 census call for the same elaborate and extensive preparations and procedures which have characterized all censuses in the U.S.S.R. since 1926, plus several innovations including the use of sampling procedures for part of the questions to be asked, the use of special schedules to collect data on the able-bodied population that is not employed and on the commuting habits of workers and students within the major urban areas, and the use of electronic computers for processing the results. This paper will describe briefly the preparations being made for the census, the enumeration procedures to be used, the questions to be asked, and the plans for processing and publishing the results.

Preparatory Work

Preparations for the census have been underway for a number of years. The Administration for the All-Union Census of Population, which was established within the Central Statistical Administration of the U.S.S.R. in 1957 to take the 1959 census, has remained as an integral part of the statistical system and presumably has been developing plans for the 1970 census from the time it completed work on the 1959 census. Since official announcement of this next census was made on May 1, 1966, there have been numerous meetings and conferences held to discuss the census program and procedures, and many articles devoted to the subject in journals and the press.

In March 1967, a test census was taken in nine rayons, or counties, of varied geographic, ecomonic, and social characteristics throughout the country. This census, which covered some 836,000 persons, was designed as a pretest of both questions to be asked and enumeration procedures. As a result of that experience, the questions contained in the present schedules were approved and numerous questions, particularly those on details of fertility, migration, and housing, were deleted. In addition, the Census Administration decided that a self-enumeration procedure did not work well and should not be used in the full census.[1]

During the latter part of 1968 and the first part of 1969, regional, county, and local offices of the Statistical Administration have been engaged in the detailed work necessary to delineate enumeration districts. Under direction from the Council of Ministers U.S.S.R., all republic, regional, and local soviets have provided assistance to the statistical offices in preparing lists of populated places, verifying boundaries of all political subdivisions, correcting updating maps, compiling lists of all prop and dwellings in each populated place, and checking the accuracy of urban dwelling registers and village household registers. On the basis of this work, which has been given great attention and apparently is being carried out in exhaustive detail, all cities and counties have been divided into census districts, instructor (crew-leader) districts, and enumeration districts.[2]

As part of this preparatory work, a mass publicity campaign has been building up, and from now until census time it will attempt to acquaint every Soviet citizen with the census and obtain his cooperation. Millions of brochures and leaflets are to be distributed, the press, radio, and television are to be utilized, and massive numbers of "agitators" from the trade unions, Komsomol, and other organizations are to give lectures and lead discussions on the census.[3]

Staffing and Training

Plans call for the recruitment of some 670,000 special personnel to conduct the census, including 540,000 enumerators, 100,000 instructors, 26,000 census district chiefs, and 4,000 assistants to the heads (inspectors) of the statistical offices at the county level.[4] This is an increase of nearly 100,000 over the 574,000 persons actually used in 1959 (of whom 468,000 were enumerators). Each enumerator in urban areas will list an average of 675 persons; each enumerator in rural areas will list an average of 575 persons.

The temporary personnel will consist of school teachers, instructors and students from colleges and vocational-technical schools, and employees of factories, offices, collective farms, and state farms. All will be selected by the local statistical office, with the approval of the county or city soviets. Employed persons will continue to receive their regular salaries while absent from their jobs, which will be from 17 working days for enumerators to 56 working days for assistants to the heads of the county statistical offices. In 1959 the temporary census workers received an enumeration bonus in addition to their regular pay, although this has not been mentioned for the 1970 census.

Brief training courses are to be held for each of the various categories of personnel. They began in June for the chiefs of the republician statistical administrations and will end in December when enumerators will be trained.

Enumeration Procedures

During a period of 4 days before the census begins, each enumerator is to make a complete tour of his district, checking the map and list of dwelling units prepared by the statistical office. He is to call on each household, verify the address and name of the head of the household, note the number of persons living there, and determine what time will be most convenient for him to call to take the census. When in the opinion of the enumerator the members of the household are capable of completing the census schedule, he will explain it to them and leave it for them to fill in by the time he returns during the census.[5]

In households where he finds able-bodied members who do not work in a factory or office or study full-time, but who work only in the household or on their small plot of land, he is to leave a special questionnaire for completion prior to his return at the time of the census. This questionnaire is designed to obtain information on the age, sex, education, specialty or profession, and desire for work of each such person and, for women in this category, the number of children and the need for their outside care. Similarly, in cities of 500,000 or more population and their adjoining counties, the enumerator is to leave at all housholds a questionnaire designed to obtain information on the so-called "shuttle" or "pendulum" migration within the urban area--that is, the age, sex, place of residence, and place of work or study of all employed persons and full-time students living in the area.

On the last day of this precensus tour, and after the enumerator has checked all dwelling units in his district, the instructor is to designate, by a "mechanical" method, every fourth unit on the enumerator's list as those units which are to be asked the sample questions. In those autonomous republics and regions established for specific nationality groups which have a population of less than 500,000, the full set of questions will be asked at every household, not every fourth, so that information in the sample schedule can be obtained and tabulated for minor territorial subdivisions.

At 8:00 a.m. on census day, Thursday, January 15, the enumerator will begin the actual enumeration. He is to visit each dwelling unit and by personal interview complete the census schedule. All persons living in the apartment or house are to be listed, by family; thus two or more families can be listed on one schedule. The enumerator also will verify and pick up any schedules or special questionnaires he left on his precensus tour.

The census is to be taken during the 8-day period from January 15 through 22. On January 22 and 23 all enumeration materials are to be turned in to the instructors for checking and verification. For the following 6 days, January 24 through 29, the instructor, with each of his enumerators, is to make a series of "control tours." In urban areas, he and the enumerator are to visit half of all dwelling units in each enumeration district, verifying all entries on the schedules for that unit, and adding any persons omitted during the census. In rural areas, the instructor is to visit all dwellings in half of his districts. Thus, all dwelling units are visited two times during the census, and half are visited three times. It is to be noted that during the control tours after the 1959 census some 285,000 additional people (0.14 percent of the total enumerated earlier) were picked up.

Population to be Covered

The census is to be taken as of a certain moment--the hour of midnight between January 14 and 15. All questions, no matter when asked, are to relate to that moment. Every person present in a dwelling unit on the census moment is to be listed, including those persons residing there temporarily. In addition, persons who usually reside in the unit but are temporarily absent are also to be enumerated. This will permit the tabulation of both an actual, or on-hand, and a usual population. Special procedures and forms are used to verify that a person temporarily absent from his usual residence is actually listed at the address given.[6] Also, special procedures will be used to enumerate the transient and institutional populations.

The population in the regions of the Far North and other areas where travel is difficult in January is to be enumerated in the fall of 1969.

Information to be Collected

There are 11 questions to be asked of 100 percent of the population:

1. Relationship to head of the family
2. For persons who are temporarily absent, but usually reside there:
   a. reason for absence
   b. length of time absent
3. For persons who are temporarily present:
   a. place of usual residence
   b. length of time absent from usual residence
4. Sex
5. Age (as of last birthday; in months for infants under 1 year of age)
6. Is person married at present time? (Yes, no, or widow-widower)
7. Nationality
8. Native language (also, no more than one other language in which fluent)
9. Education (level completed)
10. For students, indicate the type of educational institution attended
11. Souce of means of existence: work, personal agricultural plot, support by relatives, support by state (pension, stipend, children's home, etc.), other.

For the sample of 25 percent of the dwelling units, seven additional questions are to be asked of all persons who are usual residents:

12. Place of work (name of factory, organization, collective farm)
13. Occupation at place of work (for pensioners--previous occupation)
14. For a person who worked an incomplete year in 1969, record whether he worked permanently, seasonally, or temporarily, and indicate the length of time worked in 1969, in months
15. Social group (wage worker, salaried employee, collective farm member, handicraftsman, peasant with individual farm)

16. How long has person lived continuously in this populated place?
17. For a person residing here less than 2 years, indicate the place of former usual residence
18. Reason for changing place of residence.

As noted earlier, special questionnaires are also to be completed to collect information on the characteristics and circumstances of able-bodied persons who are not employed outside the household and on the characteristics and commuting habits of workers and students in all cities of 500,000 population or more.

Processing the Results

Little detailed information on the plans and procedures for processing and publishing the results of the census has been released. According to Mr. Petr Pod"yachikh, head of the Administration for the All-Union Census of Population, initial processing of the returns will be carried out at the oblast (regional) and republic offices of the statistical administration, but the final tabulation and processing will be done at the Computing Center of the Central Statistical Administration in Moscow. The schedules are not designed for use by an automatic sensing device to read and feed the information directly into a computer, although a Soviet report to the Economic Commission for Europe indicates that such a device--the "Blank" device--is to be used at some stage in the processing.[7] This report also indicates that computers of the "Minsk" series and the "Ural-14" will be used to process the census data. Pod"yachikh noted, however, that in processing the results of the test census it was learned that the system for feeding information into the computers was slow and unsatisfactory. In addition, he stated that the printing system needed improvement, and that if plans for processing and publishing the census results are to be met the equipment must be improved.[8]

Plans are to complete processing on, and presumably publish, the brief initial results of the census by the beginning of April 1970. Final processing is to be completed in the first half of 1972, and publication is to be completed in 1973. The detail in which the results are to be published is not clear, although the plans call for elaborate and detailed cross-tabulation.

Comparisons With the U.S. Census

When viewed against the content and procedures of our own 1970 census, several features of the Soviet census stand out. One is that it is a census of population only, not housing, therefore the number of questions to be asked is far less. The Soviet schedules still have fewer population items than ours, however. They contain 11 questions on a 100-percent basis and seven additional questions on a 25-percent basis; counted in the same manner, our schedules have nine questions on a 100-percent basis, 10 on a 20-percent basis, eight on a 15-percent basis, and seven on a 5-percent basis. The Soviet Statistical Administration was besieged by many requests for additional questions, particularly by academic researchers at an All-Union Conference of Statisticians in April 1968. Some requests apparently were adopted but most were turned down, primarily on the basis that the data could be better obtained through other methods--surveys or established reporting systems. The rejected proposals concerned the subjects of fertility, physical disability, the characteristics of persons in administrative positions, the presence and size of a library in the household, the number of trips taken each year and distance travelled, the presence of a radio receiver or television in the household, ethnic origin, marital status, migration, and others.[9]

In respect to procedures, Soviet census planners have now adopted the use of sampling techniques in the census, and have added the feature of using special questionnaires for certain areas or for special groups of the population. The number of temporary census personnel to be used in the U.S.S.R., 670,000 will be far larger than the 185,000 we plan to use, although the Soviet enumerators may well be employed in the census for a shorter time than our enumerators. And finally, the Soviet procedure of requiring the enumerator to call personally at all households two times, and a third time at half of the households, may indicate a more persistent effort at that level to locate all the population. Whether it truly results in a more complete and accurate census is a question that only speculation can answer. Judging from the published materials of the 1959 census, the results of the new Soviet census, whether more accurate or not, are not likely to give us the detail, either in terms of small areas or of cross-tabulation, that characterizes our own census.

[1] P. Pod"yachikh, "An Important Stage in the Preparations for the All-Union Census of Population," Vestnik statistiki (Statistical Herald), no. 10, October 1967, pp. 77-90. Some partial self enumeration procedures will be used in the census, as described in the text.

[2] V. Ivanova, "Census Regionalization--A Most Important Stage in Preparations for the Census of Population," Vestnik statistiki (Statistical Herald), no. 4, April 1969, pp. 57-64.

[3] P. Pod"yachikh, O metodologicheskikh i organizatsionnykh voprosakh Vsesoyuznoy perepisi naseleniya 1970 g. (On Methodological and Organizational Questions of the All-Union Census of Population of 1970). Materials for the All-Union Conference of Statisticians, 1968. Moscow, Statistika, 1968, pp. 36-37.

[4] Ibid., p. 31.

[5] Ibid., p. 16.

[6] "Instructions on the Conduct of the All-Union Census of Population of 1970 and the Completion of Census Schedules," Vestnik statistiki (Statistical Herald), no. 12, December 1968, pp. 45-47.

[7] Statistical Commission and Economic Commision for Europe, Conference of European Statisticians, Working Group on Electronic Data Processing. Use of Electronic Data Processing for Statistical Purposes: U.S.S.R. Report By the Central Statistical Office: New Electronic Data Processing

(EDP) Installations. Conf. Eur. Stats/WG.9/82/ Add. 20, 11 July 1968.

[8] Pod"yachikh, O metodologicheskikh, pp. 25-26.

[9] Ibid., pp. 12-15; "Readers Discuss the Questions of the All-Union Census of Population," Vestnik statistiki (Statistical Herald), no. 3, March 1968, pp. 38-44; and S. I. Bruk and V. I. Kozlov, "Ethnographic Science and the 1970 Census of Population," Sovetskaya Etnografiya (Soviet Ethnography), no. 6, June 1967, pp. 3-14.

MINUTES OF THE ANNUAL MEETING OF THE SOCIAL STATISTICS SECTION

NEW YORK, NEW YORK, AUGUST 20, 1969

The meeting was opened by Henry S. Shryock, Chairman, at 5:30 p. m. The Chairman announced that, as a result of the recent elections, the list of officers for 1970 is:

Chairman	Daniel O. Price
Chairman-Elect	Otis Dudley Duncan
Vice-Chairman	Eva Mueller
Vice-Chairman	Philip C. Sagi
Secretary	Regina Loewenstein
Section Representative on the Board of Directors	John D. Durand
Section Representative on the Council	Hubert M. Blalock
Publications Liaison Officer	Denis F. Johnston

The Chairman reappointed Edwin D. Goldfield as Editor of the Proceedings.

Margaret E. Martin, Section Representative to the Board of Directors, gave a report on the recent meeting of the Board. The Publications Committee has recommended five more editors of the Journal, and the President appointed a search committee to find more editors. The Journal is still interested in obtaining more papers on applied topics, and will have a section on Applied Papers when enough are received. Starting in December 1969, the Journal will have a section on Invited Papers.

The Publications Committee has voted to copyright the Journal of the American Statistical Association and The American Statistician. The Social Statistics Section was asked to discuss the copyrighting of the Proceedings. If the Proceedings were not copyrighted, articles that appear in the Proceedings, whether or not later in the Journal, could be reprinted in books without mention of the American Statistical Association and without approval of the authors. If there is a copyright of the Proceedings, reprinting of revised articles may need individual legal decisions.

The following motion was passed:

Future issues of the Proceedings of the Social Statistics Section should be copyrighted by the ASA, starting in 1969 if possible. When the copyright announcement is made, the American Statistical Association should issue a statement describing the reasons and the procedures for authors to obtain copyrights in their own names, if they wish to do so. A statement should also be made about the policy of allocation of fees paid for the use of copyrighted material between the authors and the American Statistical Association.

Other topics mentioned by the Section Representative to the Board were (1) group insurance, (2) plans for a Section on Computers in Statistics, and (3) plans for paid short courses to be arranged by the Section on Training. The Social Statistics Section may be asked to appoint one or more delegates to the committee to plan these short courses for members and non-members.

The Chairman read a letter from Denis Johnston, the Publications Liaison Officer. Johnston tried to get articles for the section of applied papers. After publicity, he received four articles, of which only one will probably be accepted for publication in the Journal. Further contacts of prospective authors for such papers do not seem fruitful or necessary.

Joseph Steinberg will be chairman of a committee to revise the charter of the Social Statistics Section. One of the changes will be inclusion of the Publications Liaison Officer as an officer of the section. Other necessary updating will be recommended if needed.

Suggestions for the annual meeting in 1970 were discussed. Suggested topics were criminology, especially needed criminal statistics; nutrition surveys; problems and progress of the 1970 census; trends in infant mortality rates and relationship of these rates and atomic tests in the 1950's; social networks, such as friendship structure; and international statistics, including comparisons of census methods in different countries and demographic estimates made by the United Nations.

With regard to the format of the affairs of the Social Statistics Section in 1970, some suggestions were: (1) No luncheon because of high fee, (2) coffee hour or cash bar with speaker as well as business meeting, (3) suggestions from Social Statistics Section for the general methodology lecture, (4) after the general methodology lecture, speaker to social statistics section on the applications to social statistics, with discussion by audience.

The following motion was passed:

There should be no restriction on the Program Chairman on the time or format for the section's business meeting in 1970.

1969 Officers of the Social Statistics Section

Chairman	Henry S. Shryock
Chairman-Elect	Daniel O. Price
Vice-Chairman	Eleanor B. Sheldon
Vice-Chairman	Eva Mueller
Secretary	Regina Loewenstein
Section Representative on the Board of Directors	Margaret E. Martin
Section Representative on the Council	Leslie Kish
Publications Liaison Officer	Denis F. Johnston
Editor of Proceedings	Edwin D. Goldfield