I

REPLICATION METHODS FOR STANDARD ERRORS IN COMPLEX SAMPLES

Chairman, W. EDWARDS DEMING, Consultant, Washington, D. C.

Page

BALANCED REPEATED REPLICATIONS FOR ANALYTICAL STATISTICS*

Leslie Kish and Martin Frankel, The University of Michigan

1. Summary

Balanced repeated replications (BRR) is a general method for computing standard errors. It has wide utility when specific mathematical methods are lacking, and especially for analytical statistics based on complex samples, where clustering destroys the independence of observations. We present results of methods we used since 1957 to measure standard errors of regression coefficients for several multivariate techniques. The basic design of the several samples comprised two primary selections (PS) per stratum.

Each replication was a half-sample, created by selecting one PS per stratum. The variance of the estimated coefficient b_* is measured by $(b_j-b_*)^2$ where b_j is the same estimator based on a half-sample. To increase the precision of the variance, we select repeated replications and compute the mean of the variance, $\Sigma(b_j-b_*)^2/J$. Balanced repeated replications reduce the number of repetitions needed. We obtained practically all available precision from 47 strata with 48 BRR. Though proofs are complete only for linear statistics, we offer rationale and results to indicate that BRR provides needed estimates of errors for non-linear statistics.

The ratios $\sqrt{deff}$ of actual to srs standard errors are investigated for several statistics in 5 sets of empirical studies. In each study average values of $\sqrt{deff}$ exceed 1.00, and range from 1.05 for widespread samples to 1.35 to more clustered samples.

2. Analytical statistics from complex samples

The development of standard statistical literature has been based on the assumption of samples of independent observations, which greatly facilitates obtaining interesting theoretical results. On the other hand this assumption of unrestricted random sampling is violated by the designs of most survey samples. Practical, economic designs often use clusters of sample elements, which induce strong correlations among them. These correlations have serious effects on statistics based on complex samples. They also pose formidable theoretical obstacles. It would be difficult to unravel the effects of some complex designs on the distribution of even one specific analytical statistic; it is even less reasonable to expect separate derivations for each statistic for all major designs. Hence, we need badly general methods for getting around these obstacles.

To clarify our discussion, we may think specifically of the coefficients of multiple linear regression which can typify the broader class of analytical statistics we need to estimate. The derivations of their distributions imply independence of observations (not of the variables); specific proofs of asymptotic theories, as well as laws of large numbers and central limit theorems, generally also assume independence. Yet we do conjecture that regression coefficients (and other statistics) based on large complex samples also tend in probability to approach the corresponding parameters; and that the approach is merely slowed by the correlation between elements of the same clusters. We are not alone in acting on such conjectures, and on the belief that proofs will come [Kish 1965, 2.8C; Tharakan, 1968]. After all, researchers have used data from complex samples to compute regression coefficients and other analytical statistics.

However, we should not expect that variances for regression coefficients, or other statistics, based on assumptions of independence will be valid. Rather we should expect that these will tend to underestimate the variance, as they have been shown to do for means and differences of means. We may view differences between means as the simplest form of analytical statistics, and we had abundant evidence about the design effects on their standard errors. We used these results as bases for conjectures about effects on other analytical statistics [Kish 1957; Kish 1965, 14.3]. These conjectures are born out by the results of the investigations here presented.

What alternatives have we? Research and researchers need analytical statistics, and cannot wait for the possible development of the extremely complex distribution theory necessary for complex samples. On the contrary, it is often impractical to use unrestricted random samples to conform to the distribution theory available for analytical statistics [Kish 1957, and 1965, 14.2]. Simple interpenetrating samples often will not serve because of the conflict they raise between desired stratification and adequate degrees of freedom [Jones 1956, Kish 1965, 4.4].

Many samples are highly stratified, with clustered selections from the strata. A model of two independent primary selections from each stratum is probably the most basic design that conforms adequately, if not perfectly, the actual design of many actual survey samples. Our investigations and our discussion are based on that model, and on balanced repeated replications (BRR) for obtaining estimates of standard errors, described next.

Before passing on to it we should mention a possible alternative in extensions of Taylor approximations, sometimes called "propagation of variances" or the "delta method." Tepping [1968] has recently called attention to its possibilities in complex samples, as have Deming [1960, 390 ff] and Kish [1965, 585]; it was used for standard errors of double ratios and index numbers [Kish 1968]. Note also a recent theoretical investigation [Brillinger and Tukey, 1964]. We reserve judgment until we see how useful, general,

---

* Supported by Grant GS-777 from the National Science Foundation

practical and robust it will be proven on data. Meanwhile, we suspect that BRR will fare better in most situations involving complex analytical statistics such as regression coefficients.

3. Balanced repeated replications

We propose this descriptive name for a method we used in a series of investigations since 1957, but especially since 1964. The method may be summarized in a few steps.

A. A replication (half-sample) is created by selecting at random one of the two primary selections (replicates) from each stratum. The replication reproduces accurately the complex design of the entire sample. From the j^{th} replication the desired statistics (b_{ij}) are computed. For example, compute the regression $y = b_{1j}x_1 + b_{2j}x_2 + b_{3j}x_3$ from the replication with the same estimation process used to compute $y = \hat{b}_1x_1 + \hat{b}_2x_2 + \hat{b}_3x_3$ from the entire sample. Then $(b_{1j}-\hat{b}_1)^2$ estimates the variance $\hat{b}_1$, the statistic computed from the entire sample. The variances of $\hat{b}_2$ and $\hat{b}_3$ are similarly estimated by $(b_{2j}-\hat{b}_2)^2$ and $(b_{3j}-\hat{b}_3)^2$.

Instead of $(b_{1j}-\hat{b}_1)$ we can use $(b_{1j}-b'_{1j})/2$, half of the difference between the selected replication and its complementary half-sample. The estimate $(b_{1j}-\hat{b}_1)$ is cheaper to compute, and our research shows small differences between results with the two methods. (See section 10).

B. Repeated replications are needed, because the estimates from (A), based on a single replication and a single degree of freedom, are extremely variable; almost uselessly so in most cases. But we can repeat the process (A) by drawing new replications to obtain new estimates b_{1j} and $(b_{1j}-\hat{b}_1)^2$. From $\underline{k}$ repetitions we compute the average of the $\underline{k}$ variance estimates $\sum_j(b_{1j}-\hat{b}_1)^2/k$. Because each of the $\underline{k}$ values is an estimate of the variance, so is their mean value. The precision of this average increases with the number of repetitions, but only slowly if the repeated replications are selected at random, without any connecting design. From H strata, full precision can be obtained from the 2^{H-1} possible ways of forming half-samples.

C. Balanced repeated replications reduce drastically the number of repetitions needed. For example, most of the precision from our 47 strata can be obtained from 48 balanced repetitions. This can be managed on large computers in many situations. The precision it yields is moderately adequate for usual needs: the coefficient of variation of the standard error is about $\sqrt{1/2(47)}$, or about 0.10.

D. However, a coefficient of variation of 10 percent is rather high when we want to distinguish design effects as low as 1.05 or 1.10 from 1.00. Hence we prefer to search for stability by investigating averages based on groups of statistics. A useful and common technique is to compute design effects, or deff [Kish, 1965, 812], the ratio of actual complex variances to their simple random variances. We have been averaging the values of $\sqrt{deff}$, the ratios of the standard errors, because these are less subject to extreme values. The differences between the average of deff and $\sqrt{deff}$ are not great in our practice, and the theoretical grounds for preferring one or the other not clear to us.

In computing, tabulating and accumulating design effects we have four aims: to check and improve the specific estimates of standard errors when these are highly variable; to estimate standard errors when only their simple random estimates are available; to appraise and understand sources of sampling variations; and to design better samples. We present later the results of our empirical investigations largely in terms of effects $\sqrt{deff}$, because these are more meaningful to the reader than a set of specific standard errors would be.

Simple random variances of most statistics can often be computed cheaply, based on analytical formulas already built into standard computing programs and on the entire sample; they are usually subject to much smaller sampling variations than are the actual variances computed with BRR. When analytical variances are not available, simple random variances can be computed also from simple random splits of the sample.

* * * * *

Because of its wide applicability and its essential simplicity, the method belongs perhaps to the class of "jack-knife" methods. We expect that these first applications to analytical statistics will be followed by many others, probably with modifications; we suggest some [Kish and Frankel, 1968]. That article contains results of several investigations we made to check the soundness of BRR, all very reassuring. Sections 10 and 11 contain some basic theoretical foundation and further reference.

We are aware of course that our methods, as it often happens in statistics, have run beyond rigorous mathematical foundations in places, especially in part D above. Our methods and our investigations have been stimulated by the needs of empirical research using regressions and other analytical statistics on data from complex samples. These methods are preferable to the available alternatives listed in section 2, and specifically to the universal usage of standard error estimates based on simple random assumption [Kish, 1957].

Repeated replications were first used for standard errors by the U. S. Census Bureau, as noted by Deming [1956]. At the Survey Research Center we soon began a series of computations for regression coefficients, and other analytical statistics, also introducing balancing into the repetitions. Our early programs were written by Irene Hess, Edwin Dean, and Kathleen Goode; since 1964 John Sonquist and K. S. Srikantan have contributed. Meanwhile, Margaret Gurney designed balanced replications for the U. S. Census Bureau [1963]. Walt R. Simmons used the method for surveys of the National Center for Health Statistics, and interested P. J. McCarthy, who developed an optimal method of balancing [1966, 1968]. A more efficient, flexible, and widely applicable set of programs for IBM 360 systems has been developed in 1968 by Martin Frankel and Neal Van Eck of the Survey Research Center, together with Carl Bixby and David Seigle of Interface, Inc. Ann Arbor; this work is supported by a grant from the National Center for Health Statistics.

4. Summary of empirical results

The results taken together constitute reassuring rewards for patient and difficult work since our first efforts in 1957, but especially between 1964 and 1967, when we completed the five projects described in section 5-9. These gave us some security for the following tentative conclusions. They also encouraged us to further research with better computing programs we completed in 1968.

First, the results reassure researchers who have been using data from our national samples for multivariate analyses: the standard errors computed by machine programs, based on srs assumptions, were not gross underestimates. Before these results one could reasonably fear that the design effects could be as high as they were for the standard errors of means, or conceivably even higher. (Effects of $\sqrt{deff} = 1.40$ on the standard errors of means were not rare.) For example, the average effect on standard errors was 1.17 for means, but only 1.06 for the regression coefficients from data described in section 5.

Second, design effects on standard errors of regression coefficients were shown to be estimable and of appreciable magnitudes. For example, an increase of $1.06^2 = 1.12$ in the variance corresponds to a similar decrease in the effective size of the sample. Ignoring it means that instead of error rates of 1.96 or 2.58 one is using levels of 1.85 or 2.43; hence, instead of five or one percent errors, one has 6.5 or 1.5 percent. Furthermore, the low effects of 1.06 and 1.10 were found in widespread samples; we expect greater effects in samples with greater concentrations.

These two conclusions contradict extreme positions. One extreme is the widespread wishful thinking that the design effects on the standard errors of analytical statistics, which had been generally neglected, will prove to be negligible. On the other extreme lurks the fear that those effects may eventually turn out to be as large as the large effects often found for the standard error of means; or perhaps larger. Though these extremes will remain mathematically possible for future data, their likelihood is greatly reduced by the consistency of the large and varied body of empirical data here exhibited. They show design effects consistently between the two extremes, somewhat closer to the lower. Since the design effect for means is more available, this empirical rule should be more helpful than assuming either extreme. Furthermore, the design effects for analytical statistics seem to resemble those for differences between pairs of means, also more available. For example, the mean factor of 1.06 for regression coefficients in section 5 resembles the factor of 1.07 found for differences of means for similar data.

Third, the similarity of effects on regression coefficients and on differences between means can be used. The data seem to show that, as we conjectured, the effects in regression tend to resemble the average ratios we can find more easily for differences of means for similar data. Standard errors for means are easier to compute, and we have considerable amounts of results for them. We can lean gently on those results to buttress our meager grounds for inference for regression coefficients and other analytical statistics.

Fourth, the factor $\sqrt{deff}$ seemed reasonably stable for all the coefficients in these data, across variables and equations. This is useful, because: a) individual standard errors for each variable in each equation would be expensive to compute; b) they would be subject to high variability; c) they would be difficult to present in the results [Kish, 1965, 14.1 - 14.2]. Instead for example, for the results in section 5, we rely on the factor 1.06 to adjust srs estimates of standard errors; or on $1.06^2 = 1.12$ to obtain "effective" sample sizes. Such averaging of standard errors is hazardous; of course, we investigated other methods of averaging computed standard errors; see sections 7 and 9.

Fifth, the design effect should not be assumed to be either small or uniform for all survey results. Our conjectures are to the contrary. When design effects are high for means they tend to be high for differences of means; this tendency is similar for other analytical statistics. The results in section 7 conform well to these conjectures. Differences are also found for different statistics from the same data. For example, simple correlations, and partial coefficients may have different effects than the regression coefficients.

We hope to deepen our understanding of the sources and magnitudes of these effects with research, empirical and theoretical, by others and by us. Our present phase is reminiscent of the emergence about 20 years ago of empirical results about design effects on means and aggregates.

5. Regression coefficients for a set of economic variables.

These computations, completed in 1964, were the first large scale set of results on regression coefficients. Standard errors were computed for 20 regression equations in which 7 predictor variables appear in different combinations of 1, 2, 3, or 4 at a time (see Table 5.1). They form the core of a study on Private Pensions and Individual Savings by Katona [1965, see especially Table 30].

The data represent 1,853 interviews obtained from members of the "crucial group," defined as "Complete families (husband and wife living together) with the head in the labor force and aged 35 to 64 with a family income of three thousand dollars or more." [Katona, 1965, p.8]. They constitute a subclass from a sample of 4,700 family units, selected with equal probability on three national surveys conducted in June 1962, January 1963, and June 1963. They came from the national sample of 74 primary sampling areas: the 12 largest metropolitan areas (self-representing), plus 62 other primary areas (other-representing) [Kish and Hess, 1965]. The latter were "collapsed" into 31 strata; from the former the primary units (tracts and blocks) were "combined" into 16 strata. Thus 47 strata of roughly equal size were created for the computations, in close accord with the sample design. In each of the 47 strata a pair of replicates formed the basis for the BRR computations with 48 repetitions.

The mean value of the effects $\sqrt{deff}$ for 60 regression coefficient was 1.0616. The separate values are shown in Table 5.1. When should we use the separate values and when the overall mean? Analysis of variance shows that the small differences between predictors are significant, but those between equations are not. We suggest a useful strategy: compute srs standard error and multiply by the $\sqrt{deff}$ for the predictor averaged over all equations.

Table 5.1 Effects $\sqrt{deff}$ for Standard Errors of 7 Predictors in 20 Regression Equations

Equation	1	2	3	4	5	6	7	Mean
1		1.002						1.002
2	1.030	1.006						1.018
3	1.023	1.115		1.062				1.067
4	1.172	1.011			1.016			1.066
5	1.026	1.027				0.956		1.003
6	1.090	0.970					1.194	1.085
7	1.017	1.078		1.071		0.892		1.015
8	1.070	0.978				0.929	1.149	1.032
9		1.057						1.057
10	1.141	1.048						1.095
11	1.217	0.939		1.165				1.107
12	0.997	1.109			0.966			1.024
13	1.150	1.004				1.415		1.190
14	1.083	1.077					0.986	1.049
15	1.217	0.939		1.165				1.107
16	1.109	1.052				1.096	0.971	1.057
17	1.178	1.001				1.000		1.060
18	1.186	1.068	1.015			1.165		1.109
19	0.925	0.942	0.959			1.122		0.987
20	1.258	0.985	1.029			1.082		1.089
Mean	1.105	1.020	1.001	1.116	0.991	1.073	1.075	1.062

6. Regression of voting on 4 attitude scales.

These data derive from 1,111 voters interviewed in October and again in November of 1964, in as many households selected in the 74 primary areas of the SRC national sample. A regression equation related the party receiving the vote of the respondent (as he stated it in the November interview) to 4 predictors, each of these an attitude scale of 9 points obtained in the October interviews. The sample design and the construction of 48 repetitions based on 47 computing strata were similar to those described in section 5.

Mean values of the effects $\sqrt{deff}$ on standard errors, in Table 6.1, are based on the averages of the BRR values obtained from $(b'_i-\hat{b})^2$ and $(b_i-\hat{b})^2$, the deviations of the half sample and its complement from the overall statistic $\hat{b}$. We also computed the BRR values of $(b_i-b'_i)^2/4$, and the effects from these are given in parentheses. These values are always lower, but only by a factor of 0.999 to 0.995. These small differences are reassuring and meaningful, as will be noted in section 11.

Table 6.1 Effects $\sqrt{deff}$ on Standard Errors of Regression Statistics

i	mean $\bar{x}_i$	b_i	r_{yxi}	$r_{yx_i \cdot x_j x_k x_l}$
1	1.215 (1.214)	0.958 (0.954)	1.013 (1.011)	0.935 (0.931)
2	1.018 (1.017)	1.069 (1.065)	1.104 (1.102)	1.089 (1.086)
3	1.168 (1.167)	1.136 (1.133)	1.298 (1.296)	1.198 (1.196)
4	1.021 (1.020)	0.897 (0.892)	0.970 (0.968)	0.941 (0.936)
Y	1.108 (1.107)	... ...	... ...	... ...
Mean	1.106 (1.105)	1.015 (1.011)	1.096 (1.094)	1.041 (1.037)

The mean effect 1.106 for standard errors of the 5 means is not surprising, though perhaps somewhat on the low side; 1.2 or 1.3 would be more in accord with other data of this type. This study, the smallest of the 5 we present, may be the best place to remind ourselves of the variability of our data; the coefficient of variation of these values is probably over 10 percent, $\sqrt{1/2(47)}$.

The mean effect for the multiple regression coefficient is 1.015, and this may also be on the low side. The mean effect is 1.096 for the first order correlations of the 4 predictors with the predictand, and 1.041 for the partial coefficients. The effect was 1.143 (1.138) for the multiple correlation coefficient, and 1.150 (1.146) for the adjusted coefficient.

There are 6 first order correlation coefficients between the 4 predictors, and the mean of the effects was 1.221 (1.219). (The 6 values were 1.212, 1.030, 1.320, 1.111, 1.265, 1.390.) This value is greater than we expected, and greater than the 1.106 for the means. We have no explanation for this case, and it does not hold in other projects.

To investigate the possible effects of nonnormality we repeated these computations after making a Fisher's z transformation of the coefficients. The differences for all the separate effects were reassuringly small, and the mean effect was the same to 3 decimal points.

We may add here that in our first set of BRR computations in 1957, for regression coefficients for an equation of political attitudes, we found a low mean effect, similar to the 1.015 reported here [Stokes, 1958].

7. 16 Regressions of physiological measurements of the NCHS

For 3,091 males in a national health examination survey, age, height, and weight were used as predictors of 16 physiological predictand variables, each in a separate equation. The sample came from clusters of four per segment, clustered in 42 primary areas [Simmons and Baird, 1968].

The average effect $\sqrt{deff}$ on the standard errors of means, correlations and regression coefficients is given in Table 7.1, col. (2). Individual values of $\sqrt{deff}$ for means, regression coefficients and multiple correlation coefficients are given in Table 7.2. These effects are larger than in our other results, but do not contradict them; on the contrary, this is in line with our conjectures: we expected large effects because of the large primary clusters, and because the effects on the means were large. We believe that the large effects here are not due to the nature of the variables, but to the large clusters of the sample, and possibly to the clustering of measurement errors by the survey teams.

Table 7.1 Effects of Design in 16 Regressions from 3 Predictors.

Each entry in (2) and (3) is the mean of a number in (1) values of $\sqrt{deff}$.

Statistical Type	(1)	(2)	(3)	(4) =(2)÷(3)
Ratio Means	18	1.7998	1.7549	1.0256
Simple Correlations	51	1.2616	1.2802	0.9855
Partial Correlations	48	1.3995	1.3487	1.0377
Multiple R	16	1.4653	1.4217	1.0307
Regression Coefs.	48	1.2948	1.2668	1.0221

Table 7.2 Effects, $\sqrt{deff}$, on Standard Errors of Means, Regression Statistics, and Multiple R's.

Equation	Means (Y_i)	$b(X_1)$	$b(X_2)$	$b(X_3)$	$R_{Mult.}$
1	2.022	1.126	1.169	.89	1.214
2	1.398	1.443	1.630	1.298	.942
3	1.194	1.343	1.522	1.290	.816
4	0.937	1.482	1.267	.986	1.298
5	2.093	1.160	1.441	1.514	1.543
6	1.371	1.359	1.246	1.607	1.866
7	2.023	1.219	1.426	1.564	1.701
8	1.893	1.281	1.233	1.309	1.666
9	1.885	1.361	1.465	1.110	2.036
10	1.404	1.330	1.045	1.486	1.640
11	2.470	1.550	1.374	1.323	1.836
12	1.523	0.935	1.270	1.316	1.496
13	1.833	1.178	0.857	1.147	1.175
14	2.227	1.391	1.220	1.328	2.012
15	1.204	1.453	1.355	1.165	.942
16	3.319	1.106	1.251	1.325	1.261
Mean	1.800	1.295	1.298	1.2914	1.465

These computations gave us an opportunity to investigate an issue related to the estimation scheme employed by the National Center for Health Statistics, which uses 12 age-sex categories for post-stratification. Since the post-stratification scheme is a function of the sample, the estimate based on the half-samples (i.e., the b_i and b'_i's) should be computed using post-stratification weights based on the particular half or complement half sample. Following this procedure we computed estimates of variance using a set of 16 half samples with no complements; values of $\sqrt{deff}$ for these results given in col. (2). Because the reweighting of each half sample is costly, we also computed the less costly variance estimates with weights assigned in the post-stratification of the total sample. Column (3) presents the average $\sqrt{deff}$ computed this way; for these estimates we used 24 replications with complementary half samples. Column (4) gives the ratio of the two methods; the simpler method (no reweighting of half-samples) does in general appear to underestimate the design effect. However, this degree of underestimation may perhaps be tolerated to reduce the cost of computing variances. Furthermore, to the extent that the factor of underestimation is found to be stable at 1.02 or 1.03, it may perhaps be used to adjust the cheaper estimate.

Analysis of variance of the 16 x 3 values in Table 7.2 of $\sqrt{deff}$ for the 48 regression coefficients showed no significance either for the 3 predictors or for the 16 equations. Hence for the standard error of an individual regression coefficient the best strategy may be to use 1.29 times the srs estimate of the standard error.

8. Dummy variables in regression equations

The data represents the sum of three national households samples conducted in August 1962, November 1962, and November 1963. Over 1300 interviews from each survey yielded a total of 3990, from the national sample of 74 primary areas of the Survey Research Center, similar to that in section 5. The samples are described by Lansing and Mueller in The Geographic Mobility of Labor [1967, pp 8-9, 349-358]. Again 48 balanced repetitions were used for 47 strata.

The multivariate analysis used a technique of "dummy variables" to represent nonmetric and nonscaleable predictors, and to overcome nonlinearity in other predictors. Essentially each category of every variable in the regression equation receives a value of 1 for members of the category, and 0 for nonmembers [Suits, 1957]. A standard program for the IBM 7090 (now rewritten for IBM 360) computed estimates of regression coefficients and their standard errors, as well as related statistics [Lansing and Mueller, 1967, pp 47-53, 397-417].

The program's formulas for standard errors were based, as usual, on srs assumptions. The aims of our BRR computations were to compute a set of standard errors which followed closely the complexities of the design, and especially to compute ratios of $\sqrt{deff}$ of the former to the latter. These computations were confined, due to the limitations of the program and the budget, to

6 regression equations, with a total of 64 predictor categories. See Tables 8.1. The mean value of the 64 values of $\sqrt{deff}$ is 1.10. There seems to be a fair amount of variation to be investigated later.

Table 8.1 Effects $\sqrt{deff}$ on Standard Errors of Regression Coefficient for Selected Predictor Classes in 6 "Dummy Variable" Regression Equations.

Selected Classes of Predictor Variables	Regression Equations a1	a2	b1	b2	c1	c2
College, Grad. or some	1.12	1.19	0.86	0.88	1.14	1.20
High School, Grad or some	0.86	0.80	1.02	0.75	...	...
Professional or technical	1.37	1.54	1.15	1.26	...	...
Other white collar	1.29	1.14	1.32	0.92	...	...
Blue collar	0.93	0.95	1.14	1.02	1.30	0.90
Family income > $10,000	1.10	1.44	0.98	0.86	...	...
Family income < $3,000	0.76	0.86	1.14	1.37	...	...
Financial reserves > $1,000	1.29	1.08	...	...	...	...
Financial reserves none	0.97	0.81	...	...	...	...
Unemployment, usual	1.06	1.37	1.28	0.97	...	...
Negro	0.80	0.60	0.93	0.82	0.93	0.80
Home, own or buying	1.03	1.37	0.86	0.71	1.22	1.10
Relatives, all live away	2.06	1.30	1.47	1.26	...	...
Relatives, most live away	1.15	1.20	1.07	1.04	...	...
Friends, all live away	...	...	1.46	1.23	...	...
Friends, most live away	...	...	1.45	1.07	...	...
	1.127	1.118	1.152	1.011	1.145	1.002

To conserve computational costs we selected 16 predictor classes from 9 predictor variables, as they appeared to be relevant. The 6 selected regressions equations are defined by their predictand variables, and by the subclasses on which they are based, as follows:

a) Predictand: Moved in period of one year after the study
   Subclass 1: Age under 30, n=306
   2: Age 35 and over, n=927
b) Subclass of sample: age under 35, n=979
   1) Predictand: Mobility last five years
   2) Predictand: Plans to move in next year
c) Subclass of sample: Age 35 and over, n=2991
   1) Predictand: Mobility last five years
   2) Predictand: Plans to move in next year

9. Statistics from multiple classification analysis (MCA)

A sample of 2214 family heads were interviewed in January and February 1965 in the 74 primary areas of the Center's national sample. One method for multivariate analysis consisted of an MCA equation to relate a "receptivity index" to 6 predictor variables comprising altogether 43 predictor classes; see Productive Americans by Morgan, Sirageldin, and Baerwaldt[1966, pp 360-378, 208-233].

MCA is a multivariate technique used for nonmetric data and to circumvent nonlinearity of the variables. Through iteration it obtains a least square solution for an equation relating the predictand variable to a linear expression of all the predictor classes. It is being utilized increasingly in survey research, where polytomized variables are most common. [Andrews, Morgan, Sonquist, 1967; Hess and Pillai, 1960; Kempthorne, 1952; Hill, 1959; Blau and Duncan, 1967, pp 128-140].

The wide utility and utilization of the model was the prime reason for our interest in the method. It also offered a new challenge for BRR techniques, because analytical expressions are lacking for standard errors of MCA coefficients even under srs assumptions.

Iterative methods for this large matrix were costly for our programs in 1966; our 1968 program will facilitate future computations. Hence we confined our computations to 12 partially balanced repetitions of half sample estimates of standard errors based on the complex sample design. We also wanted to compare these srs estimates in order to compute design effects; to do this we also computed 12 repetitions based on simple random splits of the sample.

One set of outputs of MCA analysis is a set of adjusted deviations; deviations for the predictand variable between the overall mean and the mean for each predictor class, after adjustment for all other variables in the equation. This deviation can also be compared to the raw unadjusted deviation for the same predictor class, thus noting the combined effect of the other variables.

For each variable MCA also yields a beta coefficient that indicates the relative explanatory value of each predictor variable; this is related to the adjusted means for all classes of that variable. Each beta may be compared to an eta for the same variable; its square is the correlation ratio which indicates the proportion of the total variance attributable to the unadjusted means of all classes of the variable.

Table 9.1 presents the computed standard errors for the eta and beta coefficients. The latter are important and seem rather stable in the neighborhood of 0.022. Both the complex and the simple random computations were based on 12 repetitions each. The latter were necessary because formulas are not available. The small number of repetitions we could afford makes these values unreliable; from the 12 strata, we expect roughly a coefficient of variation of $\sqrt{1/2(12)} = 0.2$ for the computed values. The mean values are $\sqrt{deff} = 1.222$ for the 6 values of ste(beta) and $\sqrt{deff} = 1.347$ for the corresponding ste(eta) values. We expect to check these values which appear to be higher than we expected.

Table 9.1 Standard Errors for eta and beta Coefficients in Multiple Classification Analysis. Predictand Is A Receptivity Index.

Predictor Variable	eta	ste (eta)	beta	ste (beta)
Education of Head	.5032	.0186	.1993	.0223
Age of Head	.4098	.0245	.1233	.0215
Total Family Income	.5676	.0165	.3309	.0258
Social Participation	.4296	.0258	.1586	.0232
Achievement Orient.	.3476	.0185	.1201	.0220
Sex & Marital Status	.2884	.0284	.0970	.0209

For the standard errors of the deviations a generalized table was deemed useful because of its simplicity, and because of the high variability of the 43 individual computed values. We conjectured that ste(d) = $a/\sqrt{n}$, where a is a constant and n the size of the predictor class, may be a fair approximation. For the 43 pairs of values of ste(d) and n we fitted a least square line to a log ste(d) = log a−0.5 log n. We thus obtained a value of a = 1.815 for the adjusted deviations. Similar computations for the unadjusted deviations gave a' = 2.441. The generalized table 9.2 for standard errors of deviations is given for relevant values of subclass size n. The hyperbola $a/\sqrt{n}$ appears to fit well the values of ste(d) plotted against $\sqrt{n}$; similarly for the hyperbola $a'/\sqrt{n}$.

Table 9.2 Average Standard Errors of Deviations, Adjusted and Unadjusted, As Fitted to Two Curves $a/\sqrt{n}$.

n	25	50	75	100	150	200	300	400	500	600	1640
Adjusted $1.815/\sqrt{n}$	.363	.257	.210	.182	.148	.128	.105	.091	.081	.074	.045
Unadjusted $2.441/\sqrt{n}$	.488	.345	.282	.244	.199	.173	.141	.122	.109	.100	.060

It is likely that with more data, precision, and research better approximations will be found. Although we could not discern distinct patterns for diverse predictors, they probably exist. Furthermore, the size of the sample should depend not only on n but also on other parameters which allow the design effect to vary with n, rather than to fix it with the constant a.

For our results the mean design effect was small compared to errors in measuring it; systematic variations were not detected and can be neglected. For each of the 43 deviations we took the ratio of the two computed values: the actual clustered to the srs values (each value the mean of 12 repetitions). The mean of these $\sqrt{\text{deff}}$ ratios was 1.105 for the unadjusted and 1.025 for the adjusted deviations.

To investigate sources of variability we also fitted least square lines logarithmically to ste(d) = $a/\sqrt{n}$ to estimate values of a and a' for srs estimates of ste(d), as described above for complex estimates. The estimates of a and a' are compared in Table 9.3. That table also has values of a reasonable and simple model for the curve $a/\sqrt{n}$; for unadjusted deviations this is conjectured to be $s/\sqrt{n}$ and for the unadjusted deviations $\sqrt{1-R^2}s/\sqrt{n}$. The comparison with the computed values is reasonable. We may expect a reduction from the model which, using s from the entire sample, assumes random grouping to the mean srs values for meaningful groups.

Table 9.3 Values of the Constants (a) for Three Assumptions for the Curves $a/\sqrt{n}$

	for models	for replications srs	for replications complex
Adjusted a	$s/\sqrt{n}$ = 1.769	1.725	1.815
Unadjusted a'	$\sqrt{1-R^2}\, s/\sqrt{n}$ = 2.387	2.310	2.441

10. Properties of BRR estimates

BRR techniques for estimating variances yield useful approximations for a wide variety of statistics. We present below strong justifications, we believe, for their use. We make some simplifying assumptions necessitated by the present state of the theory. A few are made here merely to keep the exposition brief. Some of these will be weakened in a fuller exposition that will also present some elaboration, and especially the results of investigations into reliability of the technique [Kish and Frankel, 1968].

For simplicity we assume here two primary selections per stratum, selected entirely independently, hence with replacement. Within each of the H strata the two primary selections are replicates of the same selection process for representing the stratum. Then from the sample S we select at random one replicate from each stratum to constitute the replication H_i. The other replicates from each stratum constitute the associated complement C_i. Note that H_i and C_i constitute two replications of the same selection process. Furthermore S also represents the same selection process but doubled in every respect.

The sample estimating function f applied to the entire sample yields the estimate $\hat{b} = f(S)$. The same function applied to replication H_i yields $b_i = f(H_i)$; applied to C_i it yields $b'_i = f(C_i)$. Our goal here is to estimate Var($\hat{b}$) by using b_i and b'_i; and to improve this estimate by repetitions of the process to constitute b_i and b'_i; etc. These may be viewed as samples from the 2^{H-1} replications that may be drawn from S.

It may help the reader if he can refer to a list of terms:

B is the population value being estimated, and we neglect differences from some true values B_{true} due to measurement errors, nonresponse, etc.

$\hat{b}$ is the statistic used by the researcher for estimating B, and variance Var($\hat{b}$) needs to be estimated; we neglect here the possible existence of some better estimator b^*.

$\bar{b}_i=(b_i+b'_i)/2$ is the mean of a replication and its complement.

$b^{(k)}=\Sigma_k b_i/k$ and $b'^{(k)}=\Sigma_k b'_i/k$ are the means of $\underline{k}$ replications and their complements, and $\bar{b}^{(k)}=(b^{(k)}+b'^{(k)})/2 = \Sigma_k \bar{b}_i/k$. The $\bar{b}^{(t)}$ would be $\bar{b}^{(k)}$ based on all of the 2^{H-1} possible replications.

$b^{(k*)}$, $b'^{(k*)}$, and $\bar{b}^{(k*)}$ denote the above based on a balanced repeated replication.

The "linear case", when the estimation function is linear in the replicate values, has notable simplicities. In the linear case $\hat{b} = \bar{b}_i$, and thus $\hat{b} = \bar{b}^{(k)}$. But this does not hold for statistics in general which are non-linear, in which our interests lie. We are chiefly interested in making estimates about $Var(\hat{b})$ from averages of $var(\bar{b}_i)$ for the non-linear cases. We also want to relate $Var(\bar{b}^{(k)})$ and $Var(\bar{b}^{(k*)})$ to $Var(\hat{b})$.

We shall denote by $E(b_i)$ the expectation of $b_i = f(H_i)$ for a specific replication H_i over all possible samples with a specified sample design combining a selection process and estimation function. Because of symmetries we have equal expectation for all $\underline{i}$ and $\underline{j}$ and their means:

$$E(b_i) = E(b'_i) = E(\bar{b}_i) = E(\bar{b}_j) = E(\bar{b}^{(k)}) = E(\bar{b}^{(k*)}) = E(\bar{b}^{(t)}). \tag{10.1}$$

However, $E(\hat{b}) \neq E(\bar{b}^{(t)})$ etc. in general.

Furthermore

$$Var(\bar{b}_i) = E\left[\frac{(b_i-b'_i)^2}{4}\right] = E(b_i-\bar{b}_i)^2 = E(b'_i-\bar{b}_i)^2, \text{ and} \tag{10.2}$$

$$Var(\bar{b}_i) = Var(\bar{b}_j) \text{ for all } \underline{i}, \underline{j}. \tag{10.3}$$

The expression $(b_i-b'_i)^2/4$ estimates the variance of $\bar{b}_i$. Although $\bar{b}_i \neq \hat{b}$ for the nonlinear case, the difference is probably often slight. However, the estimate is extremely unstable and we must obtain more precision with repetitions. If we repeat the process $\underline{k}$ times we can get the mean variance estimate $\overline{var}_k(\bar{b}_i) = \Sigma(b_i-b'_i)^2/4k$. Since the variance of all replications has the same expectation, we have

$$E\,[\overline{var}_k(\bar{b}_i)] = Var(\bar{b}_i). \tag{10.4}$$

Furthermore

$$Var(\bar{b}^{(k)}) = k^{-2}\,[\sum_i Var(\bar{b}_i) + \sum_{i\neq j} Cov(\bar{b}_i,\bar{b}_j)]$$

$$= k^{-2}\,[k\,Var(\bar{b}_i) + k(k-1)\,\rho_{\bar{b}_i,\bar{b}_j}\,Var(\bar{b}_i)]$$

$$= Var(\bar{b}_i)\left[1 - [(1-\rho_{b_i,b_j})(k-1)/k]\right] \tag{10.5}$$

where $\rho_{\bar{b}_i,\bar{b}_j}$ denotes the correlation between $\bar{b}_i$ and $\bar{b}_j$ values. For the linear case $\rho_{\bar{b}_i,\bar{b}_j} = 1$, because all $\bar{b}_i$ are equal. In our investigations we have found for orthogonally balanced replications that $\rho_{\bar{b}_i,\bar{b}_j} \geq 0.98$, when the b's are regression coefficients.

For the linear case McCarthy [1966, 1968], has shown that orthogonally balanced patterns of repeated replication produce estimates of the variance equal to the estimate that would be produced much more laboriously from all 2^H possible replications. In our investigations on nonlinear statistics we used his orthogonal balancing because we believe it to be useful. We have made investigations to relate estimates of $Var(\bar{b}_i)$ and $Var(\bar{b}^{(k*)})$ to $Var(\hat{b})$, and $\bar{b}_i$ and $\bar{b}^{(k*)}$ to $\hat{b}$. The empirical investigations were reassuring and will be presented later [Kish and Frankel, 1968]. In summary, our estimates of $Var(\bar{b}_i)$ overestimate $Var(\bar{b}^{(k)})$, and $\bar{b}^{(k)}$ is generally very close to $\hat{b}$.

We note here a justification for the use of $Var(\bar{b}_i)$ to estimate $Var(\hat{b})$ based on design effects. The half-samples H_i and C_i are constituted to preserve the full complexity of the selection process of the entire sample S. The estimation function $\underline{f}$ is the same for $\hat{b}$ as for b_i and b'_i.

From $Deff(\bar{b}_i) = Deff(\hat{b})$ we have

$$\frac{Var(\hat{b})}{\sigma^2_{\hat{b}}} = \frac{Var(\bar{b}_i)}{\sigma^2_{\bar{b}_i}} \text{ and } \frac{Var(\hat{b})}{Var(\bar{b}_i)} = \frac{\sigma^2_{\hat{b}}}{\sigma^2_{\bar{b}_i}}.$$

$\sigma^2_{\hat{b}}$ and $\sigma^2_{\bar{b}_i}$ are the variances of $\hat{b}$ and $\bar{b}_i$ under the assumption that S is a simple random sample of elements with replacement.

From nonreplicated methods we estimated $\sigma^2_{\hat{b}}$, $\sigma^2_{\bar{b}_i}$ and used their ratio as an estimate of the ratio of $Var(\hat{b})$ to $Var(\bar{b}_i)$. For the regression coefficient described in 4-9 we find that the mean value of this ratio across different regression coefficients was 1.000 with standard error (among statistics) of less than .01.

11. Approximations for $var(b^{(k)})$ and estimates of MSE from BRR

When dealing with estimation functions for linear cases we can take advantage [McCarthy 1966, 1968] of the identity between $\hat{b}$ and $\bar{b}_i$ to use the simpler computational form.

$$\overline{var}_k(\bar{b}_i) = \Sigma_k(b_i-\hat{b})^2/k = \Sigma_k(b'_i-\hat{b})^2/k = \Sigma_k(b_i-b'_i)^2/4k, \tag{11.1}$$

since $\bar{b}_i = \hat{b}$ and $(b_i-\hat{b}) = (b'_i-\hat{b})$. Here $\underline{k}$ refers to the number of computations needed to obtain BRR. Only $\underline{k}$ computations are needed to obtain the b_i values, because the b'_i values yield no new information. However, for nonlinear cases the strict equality does not hold and computing both b_i and b'_i does yield more information and improve the variance estimate. Note that

$$(b_i-\hat{b})^2 = (b_i-\bar{b}_i+\bar{b}_i-\hat{b})^2 = (b_i-\bar{b}_i)^2$$

$$+ (\bar{b}_i-\hat{b})^2 + 2(b_i-\bar{b}_i)(\bar{b}_i-\hat{b})$$

$$=(b_i-b'_i)^2/4 + e_i^2 + e_i\ (b_i-b'_i),$$

if we remember that $(b_i-b'_i) = 2(b_i-\bar{b}_i)$, where $\bar{b}_i = (b_i+b'_i)/2$ and define $e_i = (\bar{b}_i-\hat{b})$. Then averaging over $\underline{k}$ computations we have

$$\frac{1}{k}\sum_i(b_i-\hat{b})^2 - \overline{\text{var}}_k(\bar{b}_i) = \frac{1}{k}\sum_i e_i^2 + \frac{1}{k}\Sigma e_i(b_i-b'_i), \text{ and}$$

$$\frac{1}{k}\Sigma(b'_i-\hat{b})^2 - \overline{\text{var}}_k(\bar{b}_i) = \frac{1}{k}\Sigma e_i^2 - \frac{1}{k}\Sigma e_i(b_i-b'_i). \quad (11.2)$$

In the linear case both terms vanish. In our investigations the first term is always very small. Although the second terms are not large, they raise the question of strategy for averaging the two sets of variances:

$$[\sum_i(b_i-\hat{b})^2 + (b'_i-\hat{b})^2]/2k = \overline{\text{var}}_k(\bar{b}_i) + \sum_i e_i^2/k. \quad (11.3)$$

If $\hat{b}$ estimates B and if its bias is inversely related to the number of primary selections $\underline{m}$ for a specified design, then for some constant $\underline{a}$:

$$B-E(\hat{b}) = a/2m \text{ and } B-E(\bar{b}_i) = B-E(\bar{b}^k) = a/m$$

and $[E(\hat{b}-\bar{b})]^2 = (a/2m)^2 = \text{Bias}^2(\hat{b})$.

Since $E(\frac{1}{k}\Sigma e_i^2) = E[(\bar{b}_i-\hat{b})^2] \geq [E(\bar{b}_i-\hat{b})]^2$,

the use of $[\Sigma(b_i-\hat{b})^2 + (b'_i-\hat{b})^2]/2k$ will yield a conservative estimate of the mean square error $\hat{b}$.

REFERENCES

Andrews, F., Morgan, J., and Sonquist, J. [1967], Multiple Classification Analysis, Ann Arbor: Institute for Social Research

Blau, P. M. and Duncan, O. D. [1967], The American Occupational Structure, New York: John Wiley and Sons.

Brillinger, D. R., and Tukey, J. W. [1964], Asymptotic Variances, Moments, Cumulants, and Other Average Values, Princeton University: Memorandum.

Cramer, H. [1946], Mathematical Methods of Statistics, Princeton, N. J.: Princeton University Press.

Deming, W. E. [1956], "On simplification of sample design through replication with equal probabilities and without stages," JASA (51), 24-53.

Gurney, M. [1962], "The variance of the replication method for estimating variances from the CPS design," unpublished memorandum, U. S. Bureau of the Census.

Hess, I., and Pillai, R. K. [1960], Multiple Classification Analysis, Ann Arbor: Survey Research Center, Dittoed memorandum.

Hill, T. P. [1959], "An analysis of the distribution of wages and salaries in Great Britain," Econometrics, 27, 355-381.

Katona, G. [1965], Private Pensions and Individual Savings, Ann Arbor: Institute for Social Research, Monograph No. 40.

Kempthorne, O. [1952], The Design and Analysis of Experiments, New York: John Wiley and Sons (p.95).

Kendall, M. G., and Stuart A. [1958], The Advanced Theory of Statistics, Vol. I, London: Griffin and Co.

Kish, L. [1957], "Confidence intervals in complex samples," American Sociological Review (22), 154-165.

Kish, L., and Hess, I. [1959], "On variances of ratios and their differences in multi-stage samples," JASA (54), 416-446.

Kish, L., and Hess, I. [1965], The Survey Research Center's National Sample of Dwellings, Ann Arbor: Institute for Social Research.

Kish, L. [1965], Survey Sampling, New York: John Wiley and Sons.

Kish, L. [1968], "Standard errors for indexes from complex samples," JASA, (63), 512-529.

Kish, L., and Frankel, M. [1968], "Balanced repeated replications for analytical statistics," submitted to the JASA.

Lansing, J. B., and Mueller, E. [1967], The Geographic Mobility of Labor, Ann Arbor: Institute for Social Research.

McCarthy, P. J. [1966], Replication: An Approach to the Analysis of Data from Complex Surveys, Washington: Public Health Service, Series 2, No. 14.

McCarthy, P. J. [1968], "Pseudo-replication, half samples," for Symposium on Foundations of Survey Sampling, University of North Carolina, Chapel Hill.

Morgan, J. N., Sirageldin, I. A., and Baerwaldt, N. [1966], Productive Americans, Ann Arbor: Institute for Social Research.

Simmons, W. R., and Baird, J. [1968], "Use of pseudo-replication in the NCHS health examination survey," Proceedings of the Social Statistics Section of ASA.

Stokes, D. E. [1966], "Some dynamic elements of contests for the presidency," American Political Science Review, (60), 19-28.

Suits, D. B. [1957], "Dummy variables in regression equations," JASA, (55), 548-551.

Tharakan, T. C. [1968], "Inference based on complex samples from finite population, I," unfinished manuscript, Institute for Social Research, University of Michigan.

Tepping, B. J. [1968], "The estimation of variance in complex surveys," Proceedings of the Social Statistics Section of ASA.

U. S. Bureau of the Census [1963], The Current Population Survey: A Report on Methodology, Technical Paper No. 7, Washington: Superintendent of Documents.

VARIANCE ESTIMATION IN COMPLEX SURVEYS

Benjamin J. Tepping, Bureau of the Census

1. Introduction

Clearly, the variance of an estimator based on a sample survey depends upon the design of the sample as well as upon the form of the estimator. For many of the sample designs in common use, estimators of the variance of estimated totals, means, ratios and differences are readily available in the literature. For simple random samples, estimators of the variance of more complex statistics such as variances, correlation coefficients and regression coefficients are also readily available. Thus, the problem with which this session is concerned is that of estimating the variance of such statistics as compound ratios, regression coefficients, or other complicated functions of the sample observations, when the sample design is other than simple random sampling. The sample design may be multistage, the sampling units stratified at one or more levels, with probabilities of selection varying from unit to unit.

For a given statistic based on a given sample design, there will usually exist alternative estimators of the variance. The proper choice among these alternatives will be made on the basis of consideration of characteristics of their sampling distributions, such as variance, mean square error or bias, as well as on the basis of cost considerations.

In the simplest cases the variance of the statistics, for a given sample design, is a known function of certain population parameters. Those parameters may themselves be estimated from the sample, and the estimates substituted for the parameters in the variance function to obtain an estimate of the variance. For example, with a simple random sample of n units from a population of N units without replacement, the variance of the sample mean is

$$(1) \qquad \sigma_{\bar{x}}^2 = \frac{N-n}{N-1}\frac{\sigma^2}{n},$$

and the population variance σ^2 may be estimated by

$$(2) \qquad s^2 = \frac{1}{n-1}\Sigma(x_i-\bar{x})^2.$$

If, as in the case of this example, the variance function is linear in the parameters and the estimators of those parameters are unbiased, we obtain an unbiased estimator of the variance. If the function is rational, we obtain at least a consistent estimator of the variance. For example, for a simple random sample, the variance of the ratio of sample means $r = \bar{y}/\bar{x}$ is given approximately by

$$(3) \qquad \sigma_r^2 \doteq \frac{N-n}{N-1}\frac{1}{n}\frac{1}{\bar{X}^2}(\sigma_y^2+R^2\sigma_x^2-2R\sigma_{xy})$$

where $R = \bar{Y}/\bar{X}$, and $\bar{X}$, $\bar{Y}$ are the expected values of $\bar{x}$ and $\bar{y}$. One usually estimates σ_r^2 by

$$(4) \qquad s_r^2 = \frac{N-n}{N-1}\frac{1}{n}\frac{1}{\bar{x}^2}(s_y^2+r^2s_x^2-2rs_{xy})$$

where s_x^2, s_y^2 and s_{xy} are the usual estimates of the population variances and covariance.

For other statistics, or for more complex designs, the variance may not be a known function of parameters that are easily estimated for substitution into the variance formula. One may then estimate the variance by dividing the sample into random subgroups in such a way that the variance of the statistic, for a sample the size of a subgroup, can be estimated from the differences, among subgroups, of the desired statistic. If the dependence of the variance on the size and type of the subgroup is known, this leads to an estimate of the desired variance. Deming [2] has long insisted on the utility of designing the sample in such a way that the computation of such variance estimates is particularly simple and easy. While this approach has much to recommend it, it is not always appropriate.

Where it is not convenient or possible to divide the sample into sufficiently many random subgroups, the method of half-samples, termed "pseudo-replication" by McCarthy [8], has been employed by the Bureau of the Census, the Survey Research Center of the University of Michigan, and the National Center for Health Statistics and perhaps others. This method, which may be regarded as a special case of Tukey's "Jackknife" [9], involves defining subgroups which are half the size of the full sample. The subgroups are not independent but, properly constructed, lead to an estimate of the variance.

2. Half-sample estimates of variance

An attractive feature of the half-sample or pseudo-replication estimates of variance is that it is not necessary to know the exact functional form of the variance, but only the dependence of the variance on sample sizes. It should be noted that the latter cannot be taken for granted, for it is not always true that the variance is inversely proportional to sample size, although that is frequently a useful approximation.

A serious limitation to be considered is that the precision of the variance estimate depends upon the number of replications, and a sufficient number of replications may be quite costly. The Current Population Survey conducted by the Bureau of the Census is illustrative.

To simplify matters somewhat, this description* applies to the sample design and estimating procedure as used before January 1967. The sample design is multistage. The primary sampling units are large, and are classified into 357 strata, of which 112 contain only a single primary sampling unit. One primary unit is selected from each stratum with probability proportionate to its population in 1960, and a subsample of dwellings is selected in several stages in such a manner that the overall probability of selection of a dwelling is constant over the whole population of the United States.

The estimation procedure can be thought of as consisting of two successive stages of ratio estimation followed by the construction of a composite estimate. The first-stage ratio estimate applies only to those primary units in strata containing more than a single primary unit, and consists of computing an inflation factor for each of 24 groups defined by geographic region, urban-rural residence, and race, based on the characteristics of the sample primary units in 1960. The second-stage ratio estimate makes use of independent estimates of the current population by race, age, and sex to modify the inflation factors produced by the first stage.

The composite estimate reduces the variance further for many statistics by taking into account the rotation of the sample from month to month. At any given time, the sample consisted of eight subsamples, called rotation groups, of about 4,000 households each. Each rotation group itself constitutes a national sample. The rotation of the sample is such that each rotation group is retained in the sample for four successive calendar months, dropped from the sample for the next eight calendar months, and then included in the sample again for four calendar months. The rotation pattern is such that in any given month six of the rotation groups were also in the sample the preceding month. The composite estimate for month j is of the form

$$x''_j = \omega[x''_{j-1}+x'_{Ij}-x'_{I,j-1}] + (1-\omega)x'_j \tag{5}$$

where ω is a weight between 0 and 1, x'_j is the two-stage ratio estimate of the number of people having a particular characteristic based on the whole sample surveyed at time j, x'_{Ij} is the same kind of two-stage ratio estimate for month j based only on the six rotation groups that are in the sample in both months j and j-1, $x'_{I,j-1}$ is the same kind of two-stage ratio estimate for month j-1 based on the same six rotation groups, and x''_{j-1} is the composite estimate obtained for month j-1.

It can be seen that the nature of the estimator is such that a large amount of data needs to be available and processed for each replication, including data for previous months.

* For a more complete description, see [10] and [11].

McCarthy [8] has suggested a way of controlling the selection of half-samples so as to reduce the variability of the variance estimator. He has shown that, for linear statistics based on a sample of L strata, half-sample selections balanced in a certain way provide variance estimates as precise as if all possible 2^L half-samples had been used. He has also suggested the use of partially balanced replicates. From results obtained by him and by Margaret Gurney [3], [4] of the Bureau of the Census, it appears that the variability can be reduced significantly by using balanced or partially balanced replicates rather than a purely random selection of replicates. For a sample design like the Current Population Survey, the reduction in variance seems to be of the order of 1/3 when 40 partially balanced replicates are used, or about 1/6 when 20 are used. On the other hand, the variance for 40 partially balanced replicates seems to be about twice that for a completely balanced set of replicates, and the variance for 20 about four times the variance for a completely balanced set (see [4]). If these degrees of variability are not tolerable, and if larger numbers of replicates are costly, alternative ways of estimating variances are attractive.

3. Direct methods of estimating variances

The distinction between "direct" methods of estimating variances and replication methods is not always a real one. For example, in estimating the variance of the mean of a simple random sample, one may calculate the variance among the means of random subgroups (which may be the elementary units themselves) and then make use of the fact that the desired variance is a known function of the expected values of the calculated variance.

However, in the context of a complex sample design, for example a multistage sample design with two or more primary sampling units selected from each of several strata, an alternative to the use of pseudo-replication of half-samples is the following, based on an approximate linearization of the statistic involved. We may refer to equation (3), from which it is clear that

$$\sigma_r^2 \doteq \frac{N-n}{N-1}\frac{1}{n}\frac{1}{\bar{X}^2}\sigma_z^2 \tag{6}$$

if the variable z is defined by

$$z = y - Rx. \tag{7}$$

The estimation of σ_z^2 may well be easier and simpler than the estimation of σ_x^2, σ_y^2 and σ_{xy}.

This illustration is a special case of a long-known* attack on the problem of estimating a variance. The Bureau of the Census has abandoned the use of half-samples for the Current

* See for example Deming ([1], Chapter III), Kendall and Stuart ([6], Sec. 10.6), and Keyfitz [7].

Population Survey and is now estimating the variances of the quite complex composite estimator described earlier by a direct procedure. For estimating the variance of seasonally adjusted statistics, we continue to employ the replication estimator.

The direct procedure essentially amounts to calculating a linear combination of sample totals for each primary sampling unit, and then estimating the variance of the sum of those linear combinations. Thus the problem has been reduced to the simple problem of estimating the variance of a total.

The procedure may be described in the following way. Let $u = (u_1, u_2, \ldots, u_k)$ be a vector of statistics whose expected value is a vector of population parameters $U=(U_1,U_2,\ldots,U_k)$. Suppose that the population parameter of interest is a function $f(U)$, and is to be estimated by $f(u)$. To terms of the first degree in (u_i-U_i), the Taylor's series approximation for $f(u)$ is given by

$$(8) \qquad f(u) \doteq f(U) + \sum_{i=1}^{k} (u_i-U_i)\frac{\partial f(U)}{\partial U_i}$$

and hence the variance and the mean square error of $f(u)$ are, to this approximation, the same as those of the linear function

$$(9) \qquad \ell(u) = \sum_{i=1}^{k} u_i \frac{\partial f(U)}{\partial U_i}$$

where the partial derivatives are to be evaluated at $u = U$.

This "linearization" of the estimator $f(u)$ is frequently useful for the estimation of the variance of a complex estimator. The ratio estimator exhibited earlier is one example. A more complex example, the estimation of the variance of a regression coefficient based on a stratified multistage sample, can be given.

Let x_{hi}, y_{hi} denote values of the variables x, y associated with the i-th elementary sampling unit selected for the sample in stratum h. We consider the statistic

$$(10) \qquad b = \frac{\frac{1}{n}\sum_h\sum_i x_{hi}y_{hi} - (\frac{1}{n}\sum_h\sum_i x_{hi})(\frac{1}{n}\sum_h\sum_i y_{hi})}{\frac{1}{n}\sum_h\sum_i x_{hi}^2 - (\frac{1}{n}\sum_h\sum_i x_{hi})^2}$$

where $n = \sum n_h$ denotes the number of elementary units in the sample. This statistic b is sometimes taken to estimate the regression coefficient of y on x in the population, and we may be concerned in estimating its variance. Let us introduce new variables by means of the notation:

$$(11) \qquad \begin{cases} w_h = \sum_i x_{hi}y_{hi} & w = \sum_h w_h \\ u_h = \sum_i x_{hi}^2 & u = \sum_h u_h \\ x_h = \sum_i x_{hi} & x = \sum_h x_h \\ y_h = \sum_i y_{hi} & y = \sum_h y_h. \end{cases}$$

Then b may be written

$$(12) \qquad b = \frac{nw - xy}{nu - x^2}.$$

If N, W, U, X, Y denote the expectations of n, w, u, x, y respectively, the derivatives required are

$$(13) \qquad \begin{cases} \frac{\partial b}{\partial N} = \frac{X(UY-WX)}{(NU-X^2)^2} \\ \frac{\partial b}{\partial W} = \frac{N}{NU-X^2} \\ \frac{\partial b}{\partial U} = -\frac{N(NW-XY)}{(NU-X^2)^2} \\ \frac{\partial b}{\partial X} = \frac{-NUY+2NWX-X^2Y}{(NU-X^2)^2} \\ \frac{\partial b}{\partial Y} = -\frac{X}{NU-X^2}. \end{cases}$$

As before, we approximate

$$(14) \qquad Var(b) \doteq Var(n\frac{\partial b}{\partial N} + w\frac{\partial b}{\partial W} + u\frac{\partial b}{\partial U} + x\frac{\partial b}{\partial X} + y\frac{\partial b}{\partial Y}).$$

The right-hand member may be written as a sum over the strata, so that

$$(15) \qquad Var(b) \doteq Var\sum_h(\frac{\partial b}{\partial N}n_h + \frac{\partial b}{\partial W}w_h + \frac{\partial b}{\partial U}u_h + \frac{\partial b}{\partial X}x_h + \frac{\partial b}{\partial Y}y_h)$$

$$\doteq \sum_h Var(\frac{\partial b}{\partial N}n_h + \frac{\partial b}{\partial W}w_h + \frac{\partial b}{\partial U}u_h + \frac{\partial b}{\partial X}x_h + \frac{\partial b}{\partial Y}y_h)$$

since sampling is independent in the several strata. Thus the estimation of the variance of the estimator b has been reduced to the problem of estimating the variance of a linear combination of sample sums for each stratum.

The manner in which the variance of that linear combination

$$(16) \qquad \ell_h = \frac{\partial b}{\partial N}n_h + \frac{\partial b}{\partial W}w_h + \frac{\partial b}{\partial U}u_h + \frac{\partial b}{\partial X}x_h + \frac{\partial b}{\partial Y}y_h$$

may be estimated will, of course, depend upon the sample design. If, for example, the sample within each stratum is a single-stage, simple random sample of elementary units, the sample may be subdivided into, say, m equal, random subgroups. The linear expression

$$(17)\quad \ell_{hj} = \frac{\partial b}{\partial N}n_{hj} + \frac{\partial b}{\partial W}w_{hj} + \frac{\partial b}{\partial U}u_{hj} + \frac{\partial b}{\partial X}x_{hj} + \frac{\partial b}{\partial Y}y_{hj}$$

is then formed for each of the subgroups, so that

$$(18)\quad \ell_h = \sum_j \ell_{hj}$$

and the variance estimate is based on the variance among the ℓ_{hj}. On the other hand, if two or more primary sampling units were selected from each stratum and then subsampled, a quantity ℓ_{hj} may be formed for each primary unit and the variance of ℓ_h estimated from the differences among the ℓ_{hj} in precisely the same manner as the variance x_h is estimated from the differences among the x_{hj}.

One difficulty that cannot be ignored is that the coefficients in the linear form ℓ_h are unknown population parameters. The usual practice is to substitute sample estimates in the expressions for the derivatives, just as in the case of the ratio estimates one substitutes the sample ratio r = y/x for the population ratio R = Y/X in the Taylor's approximation to the variance of r. For large samples, this procedure yields satisfactory estimates. In the case of estimating the variance of the regression coefficient b, one would take

$$(19)\quad \begin{cases} \dfrac{\partial b}{\partial N} \doteq \dfrac{\bar{x}(\bar{y}-b\bar{x})}{ns_x^2} \\[2ex] \dfrac{\partial b}{\partial W} \doteq \dfrac{1}{ns_x^2} \\[2ex] \dfrac{\partial b}{\partial U} \doteq \dfrac{b}{ns_x^2} \\[2ex] \dfrac{\partial b}{\partial X} \doteq -\dfrac{\bar{y}-2b\bar{x}}{ns_x^2} \\[2ex] \dfrac{\partial b}{\partial Y} \doteq -\dfrac{\bar{x}}{ns_x^2} \end{cases}$$

where $\bar{x} = x/n$, $\bar{y} = y/n$, $\bar{u} = u/n$, $s_x^2 = \bar{u}-\bar{x}^2$.

The illustration can be extended in a straightforward way to multiple regression coefficients. If x_{ij} is taken to be the value of the i-th regressor variable for the j-th element of the sample, the estimates b_i of the regression coefficients are computed as the solution of the system of linear equations

$$(20)\quad \sum_{i=0}^{p}\sum_{j=1}^{n} x_{kj}x_{ij}b_i = \sum_{j=1}^{n} x_{kj}y_j \qquad k = 0, 1, \ldots, p$$

where $x_{oj} = 1$. We introduce new variables u_{ki} and v_k, defined by

$$(21)\quad \begin{cases} u_{ki} = \sum_{j=1}^{n} x_{kj}x_{ij} \\ v_k = \sum_{j=1}^{n} x_{kj}y_j . \end{cases}$$

The system (20) can then be written

$$(22)\quad \sum_{i=0}^{p} u_{ki}b_i = v_k \qquad k = 0, 1, \ldots, p.$$

Differentiation with respect to $u_{h\ell}$ yields

$$(23)\quad \sum_{i=0} u_{ki}\frac{\partial b_i}{\partial u_{h\ell}} = -(1-\delta_{h\ell})\delta_{kh}b_\ell - \delta_{k\ell}b_h \qquad h, k, \ell = 0, 1, \ldots, p$$

and differentiation with respect to v_h yields

$$(24)\quad \sum_{i=0}^{p} u_{ki}\frac{\partial b_i}{\partial v_k} = \delta_{hk} \qquad h, k = 0, 1, \ldots, p$$

where δ_{ij} is the Kronecker delta. The system (23) can be subdivided in $(p+1)^2$ subsystems ($h, \ell = 0, 1, \ldots, p$), each of p+1 equations in p+1 variables $\left(\frac{\partial b_i}{\partial u_{h\ell}}, i = 0, 1, \ldots, p\right)$. Actually, only $\frac{1}{2}(p+1)(p+2)$ of the subsystems are distinct because of the symmetry $u_{h\ell} = u_{\ell h}$. The system (24) can be subdivided into p+1 systems ($h = 0, 1, \ldots, p$), each of p+1 equations in p+1 variables $\left(\frac{\partial b_i}{\partial v_h}, i = 0, 1, \ldots, p\right)$. Thus the determination of the coefficients of the linear approximations to the regression coefficients will require the solution of $\frac{1}{2}(p+1)(p+4)$ systems while the half-sample method using L replications will require the solution of L systems of the same size. The amount of computation depends then on the relative values of L and $\frac{1}{2}(p+1)(p+4)$.

The limitations of this approach arise primarily from the use of the Taylor's series, for precautions must be taken to assure that the linear approximation is acceptably good. With sufficiently large sample sizes, this can usually be assured. An illustration is provided by the variance of the estimated variance for a simple random sample of size n. Here the statistic whose variance is desired is

$$(25)\quad s^2 = \frac{1}{n-1}\sum^{n}(x_i-\bar{x})^2 = \frac{1}{n-1}\Sigma x_i^2 - \frac{1}{n(n-1)}(\Sigma x_i)^2$$

If we take the variables in the estimator function f(u) to be the elementary variables x_i, then the Taylor's series is identical with the function, all of whose terms are quadratic in the x_i. Thus the linear approximation to s^2 is taken to be zero, so that the variance is estimated to be zero. The difficulty here is that the Taylor approximation is not necessarily a good one when the variables in the Taylor's series are the x_i, each of which is based on a single sample observation. But if we define new variables u and v by

$$\text{(26)} \qquad \begin{cases} u = \frac{1}{n} \Sigma x_i^2 \\ v = \frac{1}{n} \Sigma x_i \end{cases}$$

then s^2 may be written

$$\text{(27)} \qquad s^2 = \frac{n}{n-1} (u-v^2)$$

so that, evaluated at the expected values of u and v,

$$\text{(28)} \qquad \begin{cases} \frac{\partial s^2}{\partial u} = \frac{n}{n-1} \\ \frac{\partial s^2}{\partial v} = -\frac{2n\bar{X}}{n-1} . \end{cases}$$

The linear form ℓ is then

$$\text{(29)} \qquad \ell = \frac{n}{n-1} (u-2\bar{X}v)$$

whose variance is easily found to be

$$\text{(30)} \qquad \text{Var}(\ell) = \frac{n}{(n-1)^2} (\mu_4-\mu_2^2)$$

where μ_2, μ_4 denote the second and fourth moments of x. The actual variance of s^2 is

$$\text{(31)} \qquad \text{Var}(s^2) = \frac{n}{(n-1)^2} (\mu_4 - \frac{n-3}{n-1} \mu_2^2)$$

which differs only trivially from Var(ℓ) for sufficiently large n. Thus the proper choice of the variables used in this approach can be important.

4. Precision of variance estimates

It is easy to exaggerate the precision of estimates of variance for complex surveys, whether the variance is estimated by replication or by other methods, including those mentioned above. Discussions of the precision of the variance estimate usually assume that the contributions to the variance from the individual primary strata are approximately equal. Experience indicates that this is far from true. In one example (Table 1) in which there were 120 primary strata, a single stratum contributed 40 percent of the total variance between primary sampling units. Four other strata made an additional contribution of 21 percent. As a result (see Table 1), a single stratum contributed more than 80 percent of the variance of the estimated variance, and 5 of the 120 strata contributed about 95 percent of the total.

Table 2 lists the contribution of the top 5 strata to the variance between strata, the coefficient of variation of that variance estimated by the collapsed stratum method, and the percentage contribution of the top 5 strata to the variance of the estimated variance, for a number of estimates of totals and ratios. The table emphasizes the marked inequality of the contributions of the individual strata.

Tables 1 and 2 were concerned with the between-psu component of the variance. The distribution of the total variance among the strata will be somewhat less skewed, but may still be quite marked. Even in the extreme case when the within-psu variance of an estimated total is precisely the same for each of L strata, it can be shown that the proportion of the total variance contributed by stratum h is

$$\text{(32)} \qquad (\frac{1}{L} - K_h)P + K_h$$

where P is the ratio of the within-psu component of the variance to the total variance and K_h is the contribution of stratum h to the between-psu variance. Thus, for example, if the within-psu component is half the total variance, a stratum that contributed 40 percent of the between-psu variance would contribute more than 20 percent of the total variance.

Table 1: PERCENTAGE CONTRIBUTIONS TO THE BETWEEN-PSU VARIANCE AND TO THE VARIANCE OF THE ESTIMATED VARIANCE, OF THE ESTIMATED NET INCREASE OF THE U.S. POPULATION FROM MIGRATION, 1955-1960.

Stratum	Variance (%)	Variance of estimated variance (%)	Stratum	Variance (%)	Variance of estimated variance (%)	Stratum	Variance (%)	Variance of estimated variance (%)
1	0	0	41	.260	.010	81	.003	.000
2	.367	.002	42	.210	.152	82	.004	.000
3	.881	.012	43	0	0	83	.805	.130
4	.496	.014	44	.051	.000	84	.380	.007
5	.424	.009	45	.151	.002	85	.087	.000
6	.161	.003	46	0	0	86	.025	.000
7	.014	.000	47	.214	.002	87	.051	.000
8	.002	.000	48	.002	.000	88	.000	.000
9	.325	.001	49	.059	.000	89	.116	.003
10	.402	.003	50	.714	.018	90	.003	.000
11	.215	.000	51	.150	.001	91	0	0
12	.001	.000	52	.025	.000	92	0	0
13	.029	.000	53	.742	.032	93	.369	.003
14	.581	.043	54	.267	.001	94	0	0
15	.001	.000	55	.013	.000	95	.006	.000
16	.177	.000	56	.406	.001	96	1.586	1.755
17	.052	.000	57	38.986	81.883	97	.063	.001
18	.781	.011	58	0	0	98	.280	.008
19	.336	.002	59	.011	.000	99	.001	.000
20	.225	.000	60	.004	.000	100	.274	.007
21	.114	.000	61	.002	.000	101	0	0
22	.014	.000	62	.376	.048	102	2.427	.808
23	.000	.000	63	.004	.000	103	2.476	.268
24	.040	.000	64	0	0	104	.332	.002
25	.181	.001	65	0	0	105	.441	.004
26	.092	.000	66	0	0	106	.573	.004
27	3.155	.423	67	.009	.000	107	2.303	.123
28	.643	.008	68	0	0	108	1.598	.108
29	.021	.000	69	0	0	109	2.495	.144
30	.298	.003	70	0	0	110	.063	.000
31	.004	.000	71	.054	.000	111	1.624	.109
32	.245	.006	72	.702	.037	112	.772	.060
33	.056	.000	73	.050	.000	113	1.203	.030
34	.005	.000	74	.908	.089	114	.997	2.018
35	1.147	.510	75	.005	.000	115	5.731	3.563
36	.012	.000	76	.004	.000	116	.047	.000
37	.560	.023	77	.024	.000	117	.334	.011
38	.138	.002	78	.005	.000	118	.105	.005
39	.002	.000	79	5.160	1.519	119	.764	.104
40	.155	.028	80	6.882	5.665	120	2.860	.162

Each "stratum" listed above consists of a pair or a triple of non-certainty strata of the Census Bureau's Current Population Survey as used in 1966. All but 5 of the 120 groups are pairs. The estimated variance employs the collapsed group method as described in [5], Vol. I, Chapter 9, Sections 15 and 28.

Table 2: CONCENTRATION OF THE BETWEEN-PSU VARIANCE AND OF THE VARIANCE OF THE ESTIMATED VARIANCE, FOR SPECIFIED STATISTICS.

[The entries in columns (1) and (3) are the proportions (of the variances and of the variance of the estimated variance, respectively) contributed by the 5 "strata" that are the largest contributors in each case. For the definition of a "stratum" see footnote to Table 1.]

Statistic	Concentration of variance (1)	Coefficient of variation of estimated variance (2)	Concentration of variance of estimated variance (3)
Estimates of totals:			
Total population, 1960	.38	.41	.90
Rural-farm population, 1960	.37	.16	.40
Non-white population, 1960	.37	.23	.74
Live births, 1960	.31	.27	.75
Marriages, 1960	.78	1.87	.997
Number of families, 1960	.39	.36	.86
Families with 1959 income less than $3,000	.29	.15	.43
Aggregate income in 1959	.39	.29	.82
Elementary school enrollment, 1960	.31	.28	.77
High school enrollment, 1960	.29	.26	.75
Net gain through migration, 1950-1960	.60	.66	.95
Civilian labor force, 1960	.31	.31	.86
Unemployed persons, 1960	.28	.21	.63
Employed persons, 1960	.31	.31	.87
Employed in agriculture, 1960	.38	.16	.65
Employed in manufacturing, 1960	.18	.15	.38
Employed in wholesale or retail trade, 1960	.28	.19	.62
Housing units, 1960	.41	.35	.78
Vacant housing units, 1960	.68	.65	.97
Bank deposits, 1960	.31	.30	.84
Taxable payroll, January-March, 1959	.68	4.33	1.00
Value added by manufacture, 1958	.24	.14	.41
Retail sales, 1958	.23	.23	.61
Retail sales, 1954	.24	.30	.85
Wholesale sales, 1958	.46	.34	.92
Estimates of ratios to total population:			
Rural-farm population, 1960	.36	.15	.60
Non-white population, 1960	.39	.23	.74
Live births, 1960	.22	.21	.57
Marriages, 1960	.78	1.89	.997
Families with 1959 income less than $3,000	.26	.15	.46
Aggregate income in 1959	.44	.22	.85
Elementary school enrollment, 1960	.22	.21	.76
High school enrollment, 1960	.21	.22	.70
Net gain through migration, 1950-1960	.62	.68	.95
Civilian labor force, 1960	.29	.29	.84
Unemployed persons, 1960	.24	.20	.64
Employed persons, 1960	.20	.22	.65
Employed in manufacturing, 1960	.19	.14	.42
Employed in durable goods manufacturing, 1960	.26	.17	.54
Employed in wholesale or retail trade, 1960	.39	.42	.91
Housing units, 1960	.47	.48	.91
Vacant housing units, 1960	.69	.66	.97
Bank deposits, 1960	.26	.32	.82
Taxable payroll, January-March, 1959	.71	4.52	1.00
Value added by manufacture, 1958	.24	.14	.36
Retail sales, 1958	.19	.34	.85

Table 2: CONCENTRATION OF THE BETWEEN-PSU VARIANCE AND OF THE VARIANCE OF THE ESTIMATED VARIANCE, FOR SPECIFIED STATISTICS - continued.

[The entries in columns (1) and (3) are the proportions (of the variances and of the variance of the estimated variance, respectively) contributed by the 5 "strata" that are the largest contributors in each case. For the definition of a "stratum" see footnote to Table 1.]

Statistic	Concentration of variance (1)	Coefficient of variation of estimated variance (2)	Concentration of variance of estimated variance (3)
Estimates of ratios to total population continued:			
Retail sales, 1954................	.24	.39	.89
Wholesale sales, 1958.............	.43	.32	.90
Estimates of other ratios:			
Employed in agriculture/total employed, 1960..................	.37	.16	.64
Employed in durable goods/total in manufacturing, 1960..............	.17	.15	.39
Vacant/total housing units, 1960..	.70	.66	.97
Retail sales, 1958/1954...........	.47	.35	.81
Wholesale sales, 1959/1958........	.44	.34	.94

REFERENCES

[1] Deming, W. Edwards. (1943). Statistical Adjustment of Data. New York: John Wiley & Sons, Inc.

[2] Deming, W. Edwards. (1960). Sampling Design in Business Research. New York: John Wiley & Sons, Inc.

[3] Gurney, Margaret. (1964). "McCarthy's Orthogonal Replications for Estimating Variances, with Grouped Strata." Unpublished memorandum, U.S. Bureau of the Census, to be published in Technical Notes, No. 3, (1968).

[4] Gurney, Margaret. (1962). "The Variance of the Replication Method for Estimating Variances for the CPS Sample Design." Unpublished memorandum, U.S. Bureau of the Census, to be published in Technical Notes, No. 3, (1968).

[5] Hansen, M.H., W.N. Hurwitz, and W.G. Madow. (1953). Sample Survey Methods and Theory, Vol. I. New York: John Wiley & Sons, Inc.

[6] Kendall, M.G. and A. Stuart. (1958). The Advanced Theory of Statistics, Vol. I. London: Charles Griffin & Company Limited.

[7] Keyfitz, Nathan. (1957). "Estimates of Sampling Variance Where Two Units are Selected from Each Stratum." Journal of the American Statistical Association: 52, 503-510.

[8] McCarthy, Philip J. (1966). Replication: An Approach to the Analysis of Data from Complex Surveys. National Center for Health Statistics. Vital and Health Statistics, PHS Pub. No. 1000, Series 2, No. 14. Public Health Service, Washington: U.S. Government Printing Office.

[9] Tukey, J.W. (1958). "Bias and Confidence in Not-Quite Large Samples." Abstracted in Annals of Mathematical Statistics: 29, 614.

[10] U.S. Bureau of the Census. (1963). The Current Population Survey, A Report on Methodology. Technical Paper No. 7. Washington: U.S. Government Printing Office.

[11] U.S. Bureau of the Census. (1967). Concepts and Methods Used in Manpower Statistics from the Current Population Survey. Current Population Reports, Series P-23, No. 22. Washington: U.S. Government Printing Office.

PSEUDO-REPLICATION IN THE NCHS HEALTH EXAMINATION SURVEY

Walt R. Simmons and James T. Baird, Jr.
National Center for Health Statistics

1. *Introduction*

The activities which we shall describe had their origin in the question, "How should one analyze data from a complex survey?" An alternative and perhaps more specific phrasing might be "What are appropriate statistical mechanics for drawing inferences from data secured in stratified multistage probability surveys of social, demographic, or health matters?" It can be assumed that the complex survey design will be accompanied by an elaborate estimation scheme, developed within the context of finite sampling theory. The complex design and elaborate estimation generate inference problems that are different from those solved by most classical statistical analytic techniques.

Rather surprisingly, until a few years ago, this very significant issue received little attention. More recently, its importance is being recognized. See for example, Gurney (1962); McCarthy, Simmons and Losee (1965); Kish (1968); McCarthy (1968); and other references listed in the bibliography. With reluctance, we forego in the present paper any further general discussion of this fascinating question, and restrict ourselves to reporting on some practices and investigations in NCHS that are relevant to the inference problem in complex surveys. We do invite you to keep in mind, however, as we proceed, Kendall's (1961) admonition, "It will be evident that if a sample is not random and nothing precise is known about the nature of the bias operating when it was chosen, very little can be inferred from it about the parent population." And we call attention to the implication that if the variance of a sample statistic is unknown or poorly estimated, the corresponding parameter estimate may have but trivial, and at best ambiguous, value.

2. *Balanced Half-Sample Pseudo-Replication*

Before turning to an account of the way in which replication is being used in the NCHS, a very condensed synopsis is offered of theoretical work in this area carried out in the last few years by Professor Philip J. McCarthy of Cornell University, under a contractual arrangement with our Center. This work is described in greater detail in McCarthy (1966), McCarthy (1968), and in another document not yet in print.

The balanced half-sample pseudo-replication estimator of variance is described at length in an NCHS (1966) publication written by McCarthy. The essential features of the basic half-sample estimator are these:

$\bar{x}'$ is the parent sample estimator of population parameter x

$\bar{y}'_\alpha$ is an estimator of $\bar{x}$ utilizing data from only 1 of the 2 PSU's in each stratum.

$\bar{x}' = \frac{1}{2}(\bar{y}'_\alpha + \bar{y}'_\beta)$, where $\bar{y}'_\beta$ is the complement estimate to $\bar{y}'_\alpha$; i.e., is formed from all PSU's which are in the parent sample but not in the α-half-sample.

$\bar{y}'_\alpha$ and $\bar{y}'_\beta$ are statistically independent

$$E\bar{y}'_\alpha = E\bar{x}' = \bar{x}.$$

The estimator of variance is

$$s^2_{\bar{x}'} = \frac{1}{\lambda}\sum_{\alpha=1}^{\lambda}(\bar{y}'_\alpha - \bar{x}')^2$$

where λ is the number of half-sample replicates utilized.

McCarthy's research has led to a number of findings, and in particular to these conclusions.

A. For estimating the variance of a statistic which is a linear function of sample observations, it is feasible to form pseudo-replicates in such a fashion that a set of a modest number of replications produces not only an unbiased estimate of the true variance, but is identically equal to the value which would be secured if all possible pseudo-replications were formed. (For example, in a 27-strata, 2 PSU's per stratum design, a controlled set of 28 pseudo-replications yields the precise numerical result that could be obtained from the 2^{27} possible replications.)

B. The half-sample replication variance process for only the linear case is shown rigorously to be unbiased. But, very significantly, it is biased only unimportantly for a large class of statistics R (including, for example, a ratio or regression coefficient) if the expected value $E(\bar{r}-\hat{R})^2$ is satisfactorily small, where $\hat{R}$ is the parent sample estimator of R, and $\bar{r}$ is the mean value of the half-sample replicate estimators of R.

C. Estimates of variances developed by the half-sample replication method may be used in the construction of modified tests of hypotheses fully appropriate to the complex design. For example, a pseudo-Chi-square statistic can be calculated to test for independence in a two-way table of estimates, and a modified sign test, using the replicate and complement replicates, provides a low-power, but widely applicable instrument for analysis.

3. *The Health Examination Survey (HES)*

The HES is a major activity of the NCHS in which, through direct examination of a probability sample of the civilian noninstitutional population of the United States, the distributions relating to that population are secured for a considerable variety of physical and physiological characteristics, and prevalences determined for selected medical and dental conditions. The survey design incorporates such features as deep stratification, multistage clustering, controlled selection, adjustments for nonresponse, ratio estimation, and poststratification. One cycle of the survey covers a specified age range of the population and consists of examination of about seven thousand persons. The HES has been described in greater detail in several other reports, and particularly in references NCHS (1965) and NCHS (1967).

Since the cost per examined person is high in the survey, considerable effort is devoted to extracting a maximum amount of information from the data. This implies a requirement for appropriate determination of the sampling variability of a very large number of derived statistics. Computation of variance on an *ad hoc* basis by conventional techniques is practically impossible because of the extreme complexities of relevant algebraic expressions, and the large volume of computations that would be required.

4. *Estimating Variances of Aggregates and Ratios*

We describe first a standardized program in routine use in HES, whereby, using the half-sample pseudo-replication technique, variances are secured for the characteristics under study as an ancillary output to the estimates themselves. For each cell of a table, this procedure typically provides an estimate of a denominator statistic such as number of persons of a specified description, a numerator statistic such as the aggregate value of a measurement for the specified denominator class, and the ratio of numerator to denominator; it produces sampling errors for the denominator, numerator, and ratio; and it can yield for the analyst's convenience supplementary data on sample size, row percentage distributions, or age adjusted estimates.

Specifications for a particular run can be set up by an analyst in one to ten minutes. After the data have been laundered and assembled on basic tapes, running time with current programming and equipment is about 15 minutes per basic table.

An illustration of the process is given in the series of tables A-1 through A-12. These tables show the numbers of decayed, missing, and filled (DMF) teeth of adults classified by age.

Table A-1 gives the numerator under study for each cell in the table, inflated (weighted) to the estimated totals for dentulous males in the United States. For example, it shows that dentulous males age 18-24 in the United States have an estimated 15,322,000 decayed teeth—a resounding and no doubt significant statistic to dental researchers.

Table A-2 gives the corresponding weighted estimates for the denominator for each cell in the table. It shows an estimated 7,022,000 males aged 18-24. In the example this figure is the same for each column of the spread since all of the averages or ratios to be calculated have the same number of persons in the age-sex group as a common denominator. However, if the spread had been specified as some person characteristic (for example economic status), and the statistic being calculated were proportion of persons in each economic group with heart disease, say, the frequencies would, of course, vary for the spread across any given row.

Table A-1. (Numerator) Estimated DMF teeth

Sex and age	DMF	Decayed	Missing	Filled
Grand total--	1,619,629	129,096	853,831	636,719
Total, male-----	755,812	65,242	403,460	287,127
18-24 years-	93,854	15,322	32,339	46,192
25-34 years-	157,852	17,568	62,414	77,869
35-44 years-	184,498	14,042	86,714	83,741
45-54 years-	143,848	10,329	87,478	46,058
55-64 years-	99,529	5,364	71,637	22,527
65-74 years-	60,811	2,152	49,166	9,492
75-79 years-	15,416	462	13,708	1,245
Total, female---	863,816	63,854	450,370	349,591
18-24 years-	117,418	16,714	41,268	59,435
25-34 years-	185,628	17,481	78,942	89,204
35-44 years-	207,281	14,177	101,312	91,791
45-54 years-	164,654	9,609	95,959	59,085
55-64 years-	110,253	3,948	74,392	31,913
65-74 years-	66,326	1,594	48,660	16,071
75-79 years-	12,254	329	9,834	2,090

Table A-2. (Denominator) Estimated persons

Sex and age	DMF	Decayed	Missing	Filled
Grand total-	90,713	90,713	90,713	90,713
Total, male---	43,951	43,951	43,951	43,951
18-24 years-----	7,022	7,022	7,022	7,022
25-34 years-----	10,007	10,007	10,007	10,007
35-44 years-----	10,705	10,705	10,705	10,705
45-54 years-----	7,992	7,992	7,992	7,992
55-64 years-----	4,869	4,869	4,869	4,869
65-74 years-----	2,722	2,722	2,722	2,722
75-79 years-----	632	632	632	632
Total, female-	46,761	46,761	46,761	46,761
18-24 years-----	8,300	8,300	8,300	8,300
25-34 years-----	10,597	10,597	10,597	10,597
35-44 years-----	11,050	11,050	11,050	11,050
45-54 years-----	8,383	8,383	8,383	8,383
55-64 years-----	5,037	5,037	5,037	5,037
65-74 years-----	2,902	2,902	2,902	2,902
75-79 years-----	489	489	489	489

Table A-3. Average teeth per person

Sex and age	DMF	Decayed	Missing	Filled
Grand total-	17.85	1.42	9.41	7.01
Total, male---	17.19	1.48	9.17	6.53
18-24 years-----	13.36	2.18	4.60	6.57
25-34 years-----	15.77	1.75	6.23	7.78
35-44 years-----	17.23	1.31	8.10	7.82
45-54 years-----	17.99	1.29	10.94	5.76
55-64 years-----	20.44	1.10	14.71	4.62
65-74 years-----	22.33	.79	18.05	3.48
75-79 years-----	24.37	.73	21.67	1.96
Total, female-	18.47	1.36	9.63	7.47
18-24 years-----	14.14	2.01	4.97	7.16
25-34 years-----	17.51	1.64	7.44	8.41
35-44 years-----	18.75	1.28	9.16	8.30
45-54 years-----	19.64	1.14	11.44	7.04
55-64 years-----	21.88	.78	14.76	6.33
65-74 years-----	22.84	.54	16.76	5.53
75-79 years-----	25.02	.67	20.08	4.26

Table A-4. Standard deviation of numerator

Sex and age	DMF	Decayed	Missing	Filled
Grand total-	21,951	5,045	19,009	16,204
Total, male---	11,589	2,728	10,030	6,914
18-24 years-----	2,403	1,137	1,511	1,945
25-34 years-----	3,982	741	2,713	3,160
35-44 years-----	2,934	1,081	3,558	3,326
45-54 years-----	3,069	1,170	3,711	3,500
55-64 years-----	4,972	737	3,746	2,357
65-74 years-----	3,599	247	3,264	1,925
75-79 years-----	3,184	127	3,065	526
Total, female-	16,745	3,344	12,503	11,187
18-24 years-----	3,143	1,077	1,476	2,826
25-34 years-----	5,485	1,265	3,382	4,084
35-44 years-----	5,041	1,354	2,935	3,179
45-54 years-----	4,592	470	4,128	1,606
55-64 years-----	5,543	505	4,463	2,692
65-74 years-----	4,955	236	3,786	2,151
75-79 years-----	2,635	95	2,254	811

Table A-5. Standard deviation of denominator

Sex and age	DMF	Decayed	Missing	Filled
Grand total-	565	565	565	565
Total, male---	372	372	372	372
18-24 years-----	50	50	50	50
25-34 years-----	54	54	54	54
35-44 years-----	81	81	81	81
45-54 years-----	171	171	171	171
55-64 years-----	201	201	201	201
65-74 years-----	141	141	141	141
75-79 years-----	127	127	127	127
Total, female-	496	496	496	496
18-24 years-----	42	42	42	42
25-34 years-----	91	91	91	91
35-44 years-----	126	126	126	126
45-54 years-----	188	188	188	188
55-64 years-----	257	257	257	257
65-74 years-----	191	191	191	191
75-79 years-----	92	92	92	92

Table A-3 presents the calculated ratios or means which are the frequencies in Table A-1 divided by the corresponding frequencies in Table A-2. In this example the data in Table A-3 probably form the heart of the analysis from a subject matter standpoint. In other cases as, for example, the estimated proportion of persons with arthritis, the estimated total number of such persons (Table A-1—weighted numerator) would also be of interest.

Table A-4 presents the sampling errors of the estimates developed in Table A-1, i.e., the standard errors of the national estimates for the numerators.

Table A-5 contains corresponding estimates of the sampling errors of the denominator estimates shown in Table A-2.

Table A-6 presents the estimated standard error for each of the means estimated in Table A-3. This is probably the table that would be most used in evaluating observed trends in this analytical situation. Large sample normal theory would be applied for appropriate testing of hypotheses and calculation of confidence intervals.

Tables A-7 through A-12 are self-explanatory and represent auxiliary output which can be helpful to the analyst.

Table A-6. Standard deviation of ratio

Sex and age	DMF	Decayed	Missing	Filled
Grand total-	.22	.05	.20	.17
Total, male---	.21	.06	.21	.14
18-24 years-----	.33	.16	.21	.26
25-34 years-----	.38	.07	.24	.33
35-44 years-----	.24	.09	.34	.28
45-54 years-----	.36	.13	.29	.51
55-64 years-----	.46	.13	.57	.37
65-74 years-----	.45	.06	.95	.66
75-79 years-----	1.28	.19	1.60	.93
Total, female-	.27	.07	.23	.22
18-24 years-----	.35	.13	.17	.32
25-34 years-----	.52	.12	.34	.36
35-44 years-----	.37	.11	.21	.29
45-54 years-----	.36	.05	.40	.18
55-64 years-----	.42	.11	.56	.42
65-74 years-----	.46	.09	.67	.53
75-79 years-----	1.50	.17	1.99	1.38

Table A-7. Age-sex adjusted means or ratios

Sex and age	DMF	Decayed	Missing	Filled
Total-----	17.85	1.42	9.41	7.01
Male------------	17.19	1.48	9.17	6.53
Female----------	18.47	1.36	9.63	7.47

Table A-8. Numerator sample frequencies (persons)

Sex and age	DMF	Decayed	Missing	Filled
Grand total-	5,483	5,483	5,483	5,483
Total, male---	2,587	2,587	2,587	2,587
18-24 years-----	403	403	403	403
25-34 years-----	662	662	662	662
35-44 years-----	656	656	656	656
45-54 years-----	431	431	431	431
55-64 years-----	262	262	262	262
65-74 years-----	139	139	139	139
75-79 years-----	34	34	34	34
Total, female-	2,896	2,896	2,896	2,896
18-24 years-----	524	524	524	524
25-34 years-----	697	697	697	697
35-44 years-----	702	702	702	702
45-54 years-----	552	552	552	552
55-64 years-----	267	267	267	267
65-74 years-----	132	132	132	132
75-79 years-----	22	22	22	22

Table A-9. Denominator sample frequencies

Sex and age	DMF	Decayed	Missing	Filled
Grand total-	5,483	5,483	5,483	5,483
Total, male---	2,587	2,587	2,587	2,587
18-24 years-----	403	403	403	403
25-34 years-----	662	662	662	662
35-44 years-----	656	656	656	656
45-54 years-----	431	431	431	431
55-64 years-----	262	262	262	262
65-74 years-----	139	139	139	139
75-79 years-----	34	34	34	34
Total, female-	2,896	2,896	2,896	2,896
18-24 years-----	524	524	524	524
25-34 years-----	697	697	697	697
35-44 years-----	702	702	702	702
45-54 years-----	552	552	552	552
55-64 years-----	267	267	267	267
65-74 years-----	132	132	132	132
75-79 years-----	22	22	22	22

Table A-10. Rel-variance of numerator

Sex and age	DMF	Decayed	Missing	Filled
Grand total-	.00018	.00153	.00050	.00065
Total, male---	.00024	.00175	.00062	.00058
18-24 years-----	.00066	.00551	.00218	.00177
25-34 years-----	.00064	.00178	.00189	.00165
35-44 years-----	.00025	.00593	.00168	.00158
45-54 years-----	.00046	.01284	.00180	.00577
55-64 years-----	.00250	.01888	.00273	.01095
65-74 years-----	.00350	.01324	.00441	.04116
75-79 years-----	.04266	.07564	.04999	.17852
Total, female-	.00038	.00274	.00077	.00102
18-24 years-----	.00072	.00416	.00128	.00226
25-34 years-----	.00087	.00524	.00184	.00210
35-44 years-----	.00059	.00912	.00084	.00120
45-54 years-----	.00078	.00240	.00185	.00074
55-64 years-----	.00253	.01637	.00360	.00712
65-74 years-----	.00558	.02202	.00605	.01792
75-79 years-----	.04625	.08466	.05254	.15069

Table A-11. Rel-variance of denominator

Sex and age	DMF	Decayed	Missing	Filled
Grand total-	.00004	.00004	.00004	.00004
Total, male---	.00007	.00007	.00007	.00007
18-24 years-----	.00005	.00005	.00005	.00005
25-34 years-----	.00003	.00003	.00003	.00003
35-44 years-----	.00006	.00006	.00006	.00006
45-54 years-----	.00046	.00046	.00046	.00046
55-64 years-----	.00171	.00171	.00171	.00171
65-74 years-----	.00271	.00271	.00271	.00271
75-79 years-----	.04093	.04093	.04093	.04093
Total, female-	.00011	.00011	.00011	.00011
18-24 years-----	.00003	.00003	.00003	.00003
25-34 years-----	.00007	.00007	.00007	.00007
35-44 years-----	.00013	.00013	.00013	.00013
45-54 years-----	.00050	.00050	.00050	.00050
55-64 years-----	.00262	.00262	.00262	.00262
65-74 years-----	.00436	.00436	.00436	.00436
75-79 years-----	.03539	.03539	.03539	.03539

Table A-12. Rel-variance of ratio

Sex and age	DMF	Decayed	Missing	Filled
Grand total-	.00016	.00149	.00048	.00060
Total, male---	.00016	.00164	.00054	.00049
18-24 years-----	.00062	.00560	.00223	.00166
25-34 years-----	.00061	.00167	.00155	.00187
35-44 years-----	.00020	.00577	.00186	.00133
45-54 years-----	.00040	.01118	.00071	.00812
55-64 years-----	.00052	.01595	.00154	.00647
65-74 years-----	.00042	.00673	.00280	.03634
75-79 years-----	.00278	.06812	.00548	.22774
Total, female-	.00022	.00297	.00059	.00088
18-24 years-----	.00064	.00441	.00122	.00209
25-34 years-----	.00092	.00580	.00210	.00193
35-44 years-----	.00041	.00836	.00056	.00125
45-54 years-----	.00035	.00208	.00122	.00067
55-64 years-----	.00037	.02137	.00144	.00456
65-74 years-----	.00041	.02915	.00162	.00946
75-79 years-----	.00363	.06928	.00985	.10627

The tabulation described has, of course, undergone constant improvement as experience has been gained and will undoubtedly continue to do so in the future. Since the procedure was first adopted as a "production-type" program in mid-1964, several thousand such tabs have been produced. That the procedure has come to be considered a routine, convenient, and readily available analytical aid, is a positive and encouraging commentary on the applicability of computer support to statistical analysis, as well as the appropriateness and feasibility of the half-sample replication technique of variance estimation.

5. *Further Considerations in Estimating Variances*

The standard procedures just described handle very well the principal HES requirements for appropriate variances of aggregates, means, and ratios. There are many collateral problems, only a few of which can be treated here.

One group of questions relates to cost. Although we were pleased to point out that running time is only about 15 minutes for a set of tables such as A-1 through A-12, computation is still quite expensive for a large number of problems. It is believed that more powerful computers just now becoming available to the Center, along with some reprogramming, will cut costs. But it is reasonable to ask if there may not be less expensive forms of computation which are sufficiently close approximations to the full pseudo-replication technique.

Another class of problems is the extension of the pseudo-replication procedure to estimating variances of statistics other than aggregates, means, or ratios—for example, to the statistics of multiple-regression analysis.

6. *Multiple-Regression Analysis of Anthropometric Data*

In order to study both of these problems—less expensive approximating computations, and the extensions of the replication technique to other statistics—the Survey Research Center at the University of Michigan, in accordance with NCHS specifications, in a project directed by Leslie Kish, prepared for the Center an elaborate series of tables which provide methodological information about multiple regression among anthropometric measurements from the HES. These tabulations derived 16 multiple regression equations, with sampling variance being calculated for regression and correlation coefficients for each by simple random classical methods and by three versions of pseudo-replication, all carried out with three different weighting schemes (approximating devices). Again we must condense the descriptions of our investigations and offer a selection from among the findings.

Each equation studied treated one of 16 body measurements as the dependent variable in a linear regression equation with height, weight, and age as the independent, or predicting, variables. Solutions to these 16 regression problems give the statistics shown in Table B-1 as estimates of the corresponding parameters for the United States, male, noninstitutional population 18-79 years of age. The variables are identified in Table B-2.

We dispose quickly of the matter of three different versions of pseudo-replication—which are described in McCarthy (1966). These versions are labeled, "Basic," "Complement," and "Difference." Variances were calculated by all three methods, but differences among the three for a given weighting scheme are so small, that any one of the three is equally acceptable as an estimator. Indeed, if the estimators had been linear the three should have given identical results. The closeness of the three values is a comforting piece of evidence that the non-linearity in the present case is not importantly troublesome. Both this result and the validity of the pseudo-replication technique for non-linear statistics such as regression and correlation coefficients are assured if the mean of the replicate statistics is closely equal to the corresponding statistic in the parent sample. For the anthropometric data this condition was fully satisfied. For all variables and all correlation coefficients the largest mean discrepancy was that between the body measurements and age, and this was only 0.0016 with a standard deviation of 0.0014 among the 16 body measurements. Contrastingly, Table B-3 shows substantial differences between "assumed simple random" and the half-sample methods, for all weighting schemes, and this result is typical for the full set of body measurements.

Weighting Scheme I disregards the actual complex survey design and estimating procedure, and treats the data as though they were a simple random sample—i.e., each case is given unit weight.

In the Health Examination Survey three sets of adjustment factors are applied in addition to the reciprocal of the probabilities of selection, in order to take advantage of ratio estimation, poststratification, and imputation for nonresponse. Some of these calculations are made specific for defined tabulation areas in the United States, and while straightforward, involve considerable work. In the full pseudo-replication procedure these adjustment factors are calculated and applied specific for each half-sample. The process is identified here as Weighting Scheme III, or as "unique weight." When the adjustment factors of

Table B-1. Test statistics from multiple linear regression of each of 16 body measurements on age, height, and weight

Statistics	Number of tests
Mean------------------------------------	19
Simple correlation coefficient------------	51
Partial correlation coefficient-----------	48
Multiple correlation coefficient----------	16
Beta coefficient--age--------------------	16
Beta coefficient--height-----------------	16
Beta coefficient--weight-----------------	16

Table B-2. Variables used in regression analyses

Independent	Dependent
Age Height Weight	Biacromial diameter Right arm girth Chest girth Waist girth
	Right arm skinfold Infrascapular skinfold Sitting height--normal Sitting height--erect
	Knee height Popliteal height Thigh clearance height Buttock-knee length
	Buttock-popliteal length Seat breadth Elbow-elbow breadth Elbow rest height

Table B-3. Illustrative comparison of variances for alternative weighting schemes and methods of estimating variance

[Mean male chest girth measured in cm.]

Methods of estimating variance and rel-variance	Weighting Scheme		
	I Unit weight	II Con-stant weight	III Unique weight
	(Rel-variance times 10^5)		
Assumed simple random sampling---------------	.2272	.2284	.2284
Basic half-sample-------	.3625	.3771	.3256
Complement half-sample--	.3610	.3977	not cal-culated
Difference half-sample--	.3616	.3866	not cal-culated

Table B-4. Percent distribution of sample persons by relative final weight

Relative weight	Percent distri-bution
All weights-----------------------	100.0
1/2---------------------------------------	13.8
1---	58.6
2---	16.4
3---	9.3
4-7---------------------------------------	1.9

the parent or whole sample are applied unchanged to each half-sample, the process is called Weighting Scheme II, or "constant weight." This process was tested to obtain an indication as to whether the bias introduced by such a procedure is of practical significance. Weighting Scheme II involves much less expense than Weighting Scheme III, and would be preferred if it did not introduce significant bias. For the test data presented in this paper, Weighting Scheme II encompasses a fully balanced set of half-sample replications. Weighting Scheme III has incomplete balancing as a result of calculations which were not fully faithful to the design.

7. *Impact of Weighting on Estimates of Primary Statistics*

The tabulations provide empirical evidence on several aspects of estimation or weighting effects, and in particular on two classes of problems. One class consists of comparisons between the unweighted (biased) versions and the properly weighted values for primary statistics from the parent sample such as the mean, and correlation and regression coefficients. The second class consists of comparisons between variances for these same statistics computed on the one hand under the assumption of simple random design, and on the other by pseudo-replication using either Weighting Scheme II or III.

The extent of deviation from equal-weighting brought about by all steps of sampling and estimating is summarized in Table B-4 which gives the percentage distribution of sample persons by approximate size of final relative weight.

Table C-1 presents for the anthropometric data the distributions of ratios of unweighted to weighted values for the primary statistics.

The data clearly indicate, even for this design in which there is some approximation to equal weighting as indicated by the distribution in Table B-4, that numerous cases exist for which one does *not* obtain reasonably good approximations to the estimates by the use of "unweighted" data. Means were relatively stable, although a difference of one full year in mean age, for example—this was one of the observations—while not large percentagewise, is of considerable practical significance. Discrepancies larger than 10 percent were obtained for 18 percent of the linear correlation coefficients calculated. The coefficients of multiple correlation showed a possibly unexpected stability, with all "unweighted" coefficients having errors of five percent or less. Other statistics manifested expected patterns of substantial variation

Table C-1. Distributions of ratios of "unweighted" to weighted estimates, selected primary statistics, HES anthropometric data

Ratio	Mean	Correlation coefficient			Regression coefficient on:		
		Simple	Partial	Multiple	Age	Height	Weight
0.0-0.24							
0.25-0.49			1				
0.50-0.74		3	1		1		
0.75-0.89		2	4		1	1	
0.90-0.94		4	2		3	1	
0.95		3	3		1		
0.96		1	1	1		2	2
0.97	1	3	5			3	
0.98	2	6	3	3		5	3
0.99	1	4	4	4		1	1
1.00	15	10	8	7	2	1	2
1.01		5					4
1.02			1				
1.03		2	1			1	1
1.04		1	1		2		
1.05		1	2	1			1
1.06-1.10		2	4		1		
1.11-1.25		2	2		1	1	2
1.26-1.50		2	1		1		
1.51-1.75							
1.76-2.00							
Over 2.00			4		3		
Number of measures	19	51	48	16	16	16	16
Median ratio	0.999	0.994	0.995	0.995	1.00	0.977	1.00
Mean ratio	0.996	0.985	1.99	0.994	3.91	0.979	1.08
Standard deviation	0.009	0.133	5.69	0.018	9.47	0.068	0.225

between weighted and "unweighted" estimates which would tend to discourage use of the latter to approximate the former, except perhaps for an occasional preliminary or rough approximation. The causes of these differences have their source in a variety of circumstances. While we can isolate some of the factors—e.g., the difference in mean age is partly the consequence of differential nonresponse by age—the real moral to be drawn is "weight properly," and avoid the risk of unpredictable bias which may be introduced if weighting is suppressed.

8. *Variances of Primary Statistics*

The relationship of variances appropriate to the design and estimation procedure calculated by the replicate method, to those obtained by use of a simple random sampling formula have several important aspects. We shall concentrate on two of these but will note others briefly. The strategy of our approach to analysis of this matter calls for comparison of the fully balanced replication procedure with the "unweighted" simple random variances in Weighting Scheme I. We did not have the precise data preferred for this comparison. Weighting Scheme III had the full weighting methodology, but was incompletely balanced since it used a collapsed grouping of 16 strata rather than the more nearly ideal combination which treated random halves of the certainty strata as PSU's. The variances in Weighting Scheme II carried complete balancing, but used the "constant" weights as defined above, rather than the conceptually superior "unique" weights. Given this dilemma, we chose a computational course which compared first the variances under Weighting Scheme II with the simple variances of Weighting Scheme I, followed by comparison of Weighting Schemes II and III with one another.

If the ratio of replicate variance to simple random variance is near unity, the latter may serve adequately since it is much less costly, and its own sampling error is smaller. Even if the ratio were not near unity, but were approximately constant, that would permit use of the simple random sampling formula multiplied by a constant as a probably satisfactory measure of precision.

The data in Table D-1 indicate, however, not only that the simple random variances are generally too small, but that the range of ratios of proper to simple random variances show such wide ranges of variability that application of a constant "correction factor" to the unweighted simple random sample estimate of variance would not be a practical procedure for the test data studied. While there is some tendency for the ratios of correlation statistics to center about values in the range 1.5-1.8, and at 3.2 for the mean variances, the variability from one body measurement to another, is clearly too great.

There is the possibility that the unsatisfactory result displayed in Table D-1 is the consequence of *different weighting* of observations in the two estimating methods rather than in the variance formulae which distinguish between the actual design and the assumed simple random design. To explore this possibility, ratios of proper variances to simple variances were calculated, with individual observations in both numerator and denominator carrying the same weight. The distributions of ratios are shown in Table D-2. Patterns for all statistics remain essentially unchanged from those observed in Table D-1, and result in a corresponding conclusion—for the empirical situation under study a constant "correction factor" applied to the weighted simple random sampling variance estimates is not acceptable.

Table D-3 compares two sets of replicate variances, one using the constant weights of the parent sample and the other, unique weights for each replicate. The impact of this difference in weighting, when both methods employ a proper variance formula, is more difficult to evaluate. The means and medians of ratios of the two variances are near unity, and the average of the mean ratios is almost precisely unity. The variances of these ratios among the statistics are smaller than those observed in the other variance comparisons—thus indicating that the simple Weighting Scheme II might be acceptable. But the variability of the ratios still is not trivial, and on this point further calculations for a wider range of statistics are in order. Data from Tables D-1 through D-3 are recapitulated in Table D-4 in a manner which focuses attention on highlights. In condensed summary, those highlights are

1. The ordinary simple variance formula yields unsatisfactory approximations to true variance in this complex survey, whether original data are used in weighted or unweighted form. (1st two data columns, Table D-4)

Table D-1. Distributions of ratios of Weighting Scheme II replicate half-sample variance to simple random sample Weighting Scheme I variance, HES anthropometric data

Ratio	Primary statistics to which variances relate						
	Mean	Correlation coefficient			Regression coefficient on:		
		Simple	Partial	Multiple	Age	Height	Weight
0.51-1.00		5	2	2	2		1
1.01-1.50	1	19	17	4	6	6	5
1.51-2.00	2	15	14	4	6	7	7
2.01-2.50	3	7	7	1	2	2	1
2.51-3.00	2	4	4	1		1	2
3.01-3.50	4		2	2			
3.51-4.00	3						
4.01-4.50	1		1	1			
4.51-5.00	1	1	1				
5.01-5.50				1			
5.51-6.00	2						
Number of measures	19	51	48	16	16	16	16
Median ratio	3.19	1.55	1.68	1.75	1.51	1.65	1.65
Mean ratio	3.26	1.68	1.87	2.16	1.50	1.69	1.69
Standard deviation	1.22	0.689	0.805	1.24	0.433	0.428	0.527

Table D-2. Distributions of ratios of replicate variances to simple variances, all data weighted by Scheme II, selected statistics, HES anthropometric data

Ratio	Primary statistics to which variances relate						
	Mean	Correlation coefficient			Regression coefficient on:		
		Simple	Partial	Multiple	Age	Height	Weight
Under 0.51							
0.51-1.00		4	2	2	2		1
1.01-1.50	1	18	17	4	4	5	7
1.51-2.00	2	16	10	3	7	9	5
2.01-2.50	3	8	12	2	3	1	2
2.51-3.00	2	3	2	1		1	1
3.01-3.50	4	1	3	1			
3.51-4.00	3		1	2			
4.01-4.50	1						
4.51-5.00	1	1	1				
5.01-5.50	1			1			
5.51-6.00	1						
Over 6.00							
Number of measures	19	51	48	16	16	16	16
Median ratio	3.19	1.62	1.76	1.84	1.65	1.67	1.50
Mean ratio	3.23	1.73	1.90	2.19	1.60	1.69	1.60
Standard deviation	1.18	0.696	0.790	1.22	0.458	0.390	0.491

Table D-3. Distributions of ratios of replicate variance Weighting Scheme II to replicate variance Weighting Scheme III, selected statistics, HES anthropometric data

Ratio	Primary statistics to which variances relate						
	Mean	Correlation coefficient			Regression coefficient on:		
		Simple	Partial	Multiple	Age	Height	Weight
0.0-0.24			2		1		1
0.25-0.49	1				1		
0.50-0.74		7	7	5	1	2	1
0.75-0.89	1	10	15	4	4	4	4
0.90-0.94	3		4			1	4
0.95-1.05	4	7	6	1	1	4	1
1.06-1.10	3	5	3	1	2	1	3
1.11-1.25	4	7	7	2	3	2	2
1.26-1.50	1	8	2	1	2	1	
1.51-1.75	1	4	1	1	1	1	
1.76-2.00	1						
Over 2.00		3	1	1			
Number of measures	19	51	48	16	16	16	16
Median ratio	1.06	1.07	0.895	0.857	0.975	0.975	0.920
Mean ratio	1.09	1.15	0.932	1.02	0.966	0.995	0.896
Standard deviation	0.304	0.473	0.327	0.472	0.373	0.255	0.248

Table D-4. Recap of highlights from Tables D-1 through D-3

Type of statistic	Median ratios of estimates of variance		Standard deviations of ratios of variances over all test variables		
	Correct to unweighted simple random variances	Correct to weighted simple variances	Correct to unweighted simple random variances	Correct to weighted simple variances	Replicate Weighting Scheme II to replicate Weighting Scheme III[1]
Mean	3.19	3.19	1.22	1.18	0.30
Correlation coefficient					
Simple	1.55	1.62	0.69	0.70	0.47
Partial	1.68	1.76	0.80	0.79	0.33
Multiple	1.75	1.84	1.24	1.22	0.47
Regression coefficient					
Age	1.51	1.65	0.43	0.46	0.37
Height	1.65	1.67	0.43	0.39	0.26
Weight	1.65	1.50	0.53	0.49	0.25

[1]The overall average ratio for this comparison is 1.007.

2. Even multiplication of simple variances by a calibrating constant would not produce adequate measures of sampling variance. (Note the substantial variations shown in 3rd and 4th data columns of Table D-4.)

3. In both the above comparisons, whether the data are weighted or unweighted makes little difference in the estimated variance.

4. When the pseudo-replication method is employed, it is unclear whether the "constant" weights of Scheme II are satisfactory approximations for the "unique" weights of Scheme III. The simpler scheme appears to be unbiased, but may understate or overstate variance too frequently. (Data column 5 and the footnote of Table D-4)

9. *Addenda*

9.1 *Addendum I.* NCHS at present has in routine use only the programs described earlier for handling aggregates and means or ratios, but there should be available in the fairly near future a pseudo-replication multiple regression analysis package and several other statistical measures which are being programmed by SRC, Michigan, under contract with the Center.

9.2 *Addendum II.* It may be that the most significant feature of the pseudo-replication process has not yet been exploited. Consider once more the 2-way table of DMF teeth. As a step in the process of calculating variances, the program builds quite literally 28 half-sample replicates of that parent table. Each of these half-sample replicates is an estimator of the same statistics as appear in the parent table. The variance among the replicates *is* the variance of their mean value. At the analyst's option each of the replicates can be printed out. When that is done, there is available not just a mean and a variance, but a *distribution* of 28 correlated estimates. Indeed the analyst now has before him 28 pseudo-replications of an experiment, as it were. Thus far, we have used this wealth of material only occasionally, and in unstructured ways. (The print-outs have been used also as the vehicle for calculating variance of medians and other position statistics.) But we suspect that the replicated half-sample tables contain building blocks that can become the basis for significant analytical structures of a new order. We commend study of this intriguing possibility to researchers.

REFERENCES

Gurney, Margaret, "The Variance of the Replication Method for Estimating Variances" (Unpublished memorandum, U.S. Bureau of the Census, Washington, D.C., 1962).

Kendall, M. G., and Stuart, A., *The Advanced Theory of Statistics,* Vol. II, p. 1, Hafner Publishing Co., New York, 1961.

Kish, Leslie, "Design and Estimation for Subclasses, Comparisons, and Analytical Statistics," paper presented at the Symposium on Foundations of Survey Sampling, Chapel Hill, N.C., April 22-26, 1968.

McCarthy, Philip J., Simmons, Walt R., and Losee, Garrie J., "Replication Techniques for Estimating Variances from Complex Surveys," paper presented at Joint Session of the Epidemiology and Statistics Sections, APHA, Chicago, Ill., October 18, 1965.

McCarthy, Philip J., *Replication: An Approach to the Analysis of Data from Complex Surveys,* PHS Publication No. 1000, Series 2, No. 14, Washington, D.C., April 1966.

McCarthy, Philip J., "Pseudo-Replication: Half-Samples," paper presented at the Symposium on Foundations of Survey Sampling, Chapel Hill, N.C., April 22-26, 1968.

National Center for Health Statistics, *Plan and Initial Program of the Health Examination Survey,* PHS Publication No. 1000, Series 1, No. 4, Washington, D.C., July 1965.

National Center for Health Statistics, *Plan, Operation, and Response Results of a Program of Children's Examinations,* PHS Publication No. 1000, Series 1, No. 5, Washington, D.C., October 1967.

DISCUSSION

Margaret Gurney, Bureau of the Census

My comments have to do chiefly with the reliability of estimates of the variance made from the sample, whether the method used is the collapsed stratum method, the replication method, or the "direct" method described by Dr. Tepping.

The paper by Kish and Frankel uses $\sqrt{1/2L}$ as a measure of the coefficient of variation of the standard error of an estimate from a design with L = 47. This result is obtained if we assume that β is approximately 3, in the well-known formula for the rel-variance of the estimate of the variance [1].

The assumption that β is about 3 may be far from true, as is indicated in Tables 1 and 2 of the Tepping paper: much of the variance, and substantially more of the variance of the variance (or the CV of the standard error) may come from a few (sometimes only 1) of the strata. In examining the between stratum variances for several of the characteristics in his tables we found values of β of 33, 68, and 157, for individual strata. Combining 2 strata (as is done in the collapsed stratum technique, which is used in all of these papers) may result in an average value of β which is nearly as large as the larger of the β's for the two strata, if the pairing of strata for collapsing has been inefficient for the particular estimate being considered.

If, as implied in the Tepping Tables 1 and 2, a few strata may dominate not only the variance, but also the variance of the estimated variance, the average β for the whole sample design may be much larger than 3. If, for example, the average β is of the order of 33, the formula for the CV of the standard error with the CPS design and 120 paired strata would lead to a CV of more than 25 percent; with fewer strata, such as 47 pairs, the CV would be larger, about 40 percent.

Admittedly, a β of 33 for the whole sample design seems quite large, but there are many important agricultural crops (for example rice, sugar cane, citrus fruits) which are grown in only a few parts of the country, and for these β may be appreciably greater than 3. Similarly, there are many important industries which are localized, and for which a national sample might produce β's which are larger than 3.

It is important, therefore, to know what is going into the variance. With the replication method this is difficult, since we do not see the individual original strata. If most of the variance comes from one collapsed stratum, one-half of the balanced replication estimates may be much larger than the other half; if most of it comes from two collapsed strata one-fourth of the replicates may be inordinately large. But the individual contributions to the estimate of the variance are not displayed.

This discussion of the distribution of the replication estimates leads to mention of Addendum II on page 34 of the Simmons-Baird paper, where it is suggested that much could be done with the individual estimates which are produced by the replication method (28 estimates in the HEW survey of the paper). The idea of displaying these individual estimates as a routine part of production of the data is a good one -- it is not new, having been stressed on numerous occasions by the chairman of our meeting. It brings the analyst one step closer to the original data: it can be used to some extent as a quality control device, in that finding an estimate which deviates considerably from the other estimates may indicate that something has happened in the processing of that replicate. The distribution of the estimates gives some feeling for the variability of the replication estimates, and may indicate that most of the variability comes from one or two strata. It could perhaps take the place of variance calculations, for less important or infrequently collected statistics.

[1] Hansen, M.H., W.N. Hurwitz, and W.G. Madow. (1953). Sample Survey Methods and Theory, Vol. I, p. 427. New York: John Wiley & Sons, Inc.

Discussion

John W. Tukey, Princeton University

If I might start with a light touch, I would point out that by selecting the last letter of the word "effect" rather than the third letter of that word, as used by Kish and Frankel, we would get DEFT rather than DEFF. The implications of the modified term seem quite appropriate.

In dealing with advantages of half-sample, jackknife, and related methods, we should not neglect the advantages of these methods in showing the client actual examples of varying numbers. This can give the naïve client a much better grasp of the true situation. (This point is made in the abstract of the paper in Session 63 by Mr. Gedanken.)

One important point about the use of the jackknife--in which, rather than leaving out half of the available data, one leaves out smaller pieces in turn until all has been left out once--is its ability to be used at two or more levels. If one had used the jackknife method rather than the half-sample method to obtain the DEFF or DEFT values, as in Kish and Frankel's situation, one could go ahead to estimate the stability of these results, or of their differences, or their ratios. By doing this we would have a better understanding of what these results, as well as many others, really mean.

The technique is simple in principle, but often not easily grasped without detailed exposition. The basic idea in dealing with a DEFT, for example, would be to lay aside one piece of the data and then calculate the DEFT by jackknifing the remainder. This jackknifing would involve leaving out additional pieces of the data, one at a time, and in turn. Once this has been done for one first-stage piece, we proceed to do all this over again and again, laying aside each piece of the data at the first stage in turn. We now have one value of DEFT corresponding to the laying aside of each piece at the first stage. Once we have repeated all this once more, leaving out nothing at the first stage, we are ready to jackknife the DEFTs thus obtained and thus estimate their variability. (For a more detailed account of this general sort of procedure, see the chapter by Mosteller and myself "Data Analysis, Including Statistics" which will appear in the next few weeks in the second volume of the second edition of the HANDBOOK OF SOCIAL PSYCHOLOGY edited by Gardner Lindzey and Elliot Aronson and published by Addison-Wesley.)

It seems to me that there will, in the two-psu-per-stratum situation faced by Kish and Frankel, prove to be real advantages to a suitable jackknife procedure--one that leaves out more than one psu, but less than half of all psu's. If we have five strata, each with two psu's, the half-sample method requires leaving out one psu in each strata, which can be done in 32 ways. A probably sensible jackknife approach would involve leaving out one psu in each of, say, two of the five strata. There are 40 possible ways to do this. The gain will come from leaving out enough, but noticeably less than half of the data.

Some work on the jackknife in situations with two-way classification, situations similar but not equivalent to this two-psu-per-stratum case, is included in an unpublished Princeton thesis by Donald Burdick. Extensions to the situation just mentioned should also, it would seem, proceed smoothly.

A simple example on which to compare "jackknifing" and "halving" is the problem of data gathered in several blocks with three values, equally spaced in time, obtained in each block. This sort of data arises naturally in many agricultural problems (including time of planting and time of harvesting). Yates (private communication) suggested that, where the number of blocks was a power of two, we treat such situations by halving the data and comparing the halves, repeating this in an interesting and ingenious way according to a fractional factorial pattern, thus obtaining the full number of degrees of freedom for the variability estimate.

Analysis of this problem shows that the bias due to halving--both in the location of the optimum data and in the estimate of the variance of this optimum data--is noticeably larger for halving than for "leaving out one" jackknifing, which also provides the full number of degrees of freedom for a variance estimate. I believe we can expect to find this phenomenon rather general. Accordingly, I believe that "leave out a few" techniques will do even better than halving in the two-psu-per-stratum situation.

I am pleased to see the interest in, and active use of all these methods. I am sure we will see much more of it.

II

SURVEY ANALYSIS: USES OF CLUSTER AND PATTERN RECOGNITION TECHNIQUES

Chairman, WILLIAM G. MADOW, Stanford Research Institute

ON THE STATUS OF APPLICATIONS OF CLUSTERING TECHNIQUES TO BEHAVIORAL SCIENCES DATA

Geoffrey H. Ball, Stanford Research Institute

Herman P. Friedman, International Business Machines Corporation

1.0 Introduction

What is clustering? One definition is provided by Professor David Wallace,

> "Clustering methods form a loosely organized body of techniques for the analysis of [multivariate] data. As with most methods of data analysis the aim is to find, describe, and hopefully to lead to an explanation of simple structure in a complex mass of data. Clustering methods are distinguished by the type of structure that is sought." D. Wallace (1968)

What kind of structure is sought by clustering techniques? It might be data having a structure that can be usefully described as a mixture of normal distributions with covariance matrices that are:

(a) all the identity matrix

(b) all equal but not equal to the identity matrix

(c) different for each of the underlying distributions.

For unequal covariance matrices, the data might take on the appearance shown in Figures 1 and 2. (Such highly artificial data sets are useful because they provide convenient methods of testing the characteristics of an existing or proposed clustering technique prior to its use on data having unknown characteristics). More generally distributions are not normal, as for example, in the data describing luminosity as a function of star temperature shown in Figure 3.

In the two and three dimensional data that we have shown this far, it is possible to plot the data, to see its structure and from this visual observation to reach conclusions regarding its cluster structure. (However, mixtures of even straightforward multivariate normal distributions do no easily reveal their structure when the variances are so large that the distributions overlap). In p dimensions it is not possible to see the p-dimensional relationships, and it is for this reason that the data analyst must rely on techniques such as clustering to find the structure implied by the numeric values of the data samples.

There are a large body of "clustering" procedures that take as their basic data a matrix of distances or similarities between each pair of objects. A basic problem here is the appropriate definition of this measure of similarity. This is particularly significant since the choice of distance implicitly defines one aspect of the structure. Here the work of multidimensional scaling as exemplified by the works of Shepard (1962), Kruskal (1968), Torgerson (1968) and Guttman (1968) are relevant. Their work enables one to take non-metric data for which similarities or distances have been specified and produces a metric space of quantitative dimensions in which the clustering procedures may search for group structure. The interplay between these two techniques is beginning to be recognized and is discussed by Torgerson (1968) and Green et al (1968).

Actually we believe there is an inherent circularity in talking about finding structure in data that is brought about by a looseness in language. It appears to us that the questions to be answered are akin to those in using models in science. We no longer ask for the correct (absolute) model. We do ask when is a model better than another with respect to powers of prediction, explanation, and simplicity. In the same way we look at models of data structures and ask when does one model fit the data better than another.

The choice of one model of data structure over another may depend on how well the results can be interpreted within the paradigm of the subject matter that gave rise to the data. We thus beg the question of what is "structure" in the data and deal with the question of when does one model fit the data better than another. This is easier said than done in general in the field of data analysis and in cluster analysis in particular.

Although a solid, coherent foundation of cluster analysis and pattern recognition does not yet exist, a wide variety of different procedures have been applied to the problem of grouping data.

2.0 Applications of Cluster Analysis

We list here some of the diverse disciplines that have used clustering techniques:

> <u>Geography</u>--Regions in which various geographic or demographic characteristics are relatively similar are useful constructs for the geographer and have been found by clustering techniques. (Berry, 1960)
>
> <u>Economics</u>--For aggregation of various kinds of industries into groups having similar characteristics with regard to the economic analysis to be performed. (Fisher, 1958)
>
> <u>Electrical Engineering</u>--For the

detection of signals of unknown characteristics that recur frequently in a background of random noise. (Fralick, 1967)

Information Retrieval--To find classes of descriptors for articles and papers. (Dale and Dale, 1965)

Medicine--To group electrical cardiograms into subgroups. (Stark et al, 1962)

Numerical Taxonomy--To group species of living organisms into hierarchic trees by an explicit mathematically defined method to contrast the trees obtained with those obtained by older methods that contained considerable implicit judgement on the part of the taxonomist. (Sokal and Sneath, 1963)

Psychology and Sociology--To group people into types that may relate to treatment categories or behavior categories. (Tryon, 1967)

Statistics--For obtaining minimum variance stratified sampling partitions of a range of a variable. (Dalenius, 1951)

From this list it is clear that there has been a considerable interest in many disciplines in techniques that are able to take vectors and group them into subgroups in a way that makes some intuitive sense and that is useful in organizing data related to a variety of subject matter areas.

The need has always been around to group multivariate data, and some of the previous examples such as taxonomies of psychiatric groupings indicate that there was a hope that the groupings would have some organizing, some explanatory and some predictive power, for example, in taxonomy that we would be able to understand the evolutionary lineage of various organisms, or in the case of psychiatric groups to indicate treatment classes, where the same treatment could be used on a particular group.

The reader interested in pursuing further the work of cluster analysis in the field of psychiatry would be well advised to look at Katz, Cole, Barton (1968)---"The Role and Methodology of Classification in Psychiatry and Psychopathology", where he will find many points of view including those of clinicians as well as data analysts. In numerical taxonomy, the book by Sokal and Sneath (1965) remains the standard reference. In the area of market analysis the monograph by Green et al (1968) is one of the most complete we are aware of. The paper by Ball (1965) contains excellent references for the person interested to see the wide variety of different methods that were developed in numerous disciplines.

3.0 Clustering as a Data Analysis Tool

We see cluster analysis as one useful tool in the highly iterative process of data analysis. Clustering can be put to the following uses:

(a) It can suggest, by grouping the data into groups with high intra-group similarity, multiple working hypotheses that are appropriately vague and hence appropriately suggestive of alternate views of the data. It appears to us that this grouping is an important aspect of concept formation and of theory development, since in both cases grouping of things sufficiently similar to be considered as a unitary item can lead to important advances in understanding.

(b) Clustering provides a way of starting an analysis, and it can provide a flexible set of initial categories for further modification.

(c) It suggests possible ways of decomposing the data into simpler subsets of objects or variables that can be examined graphically, so that the analyst obtains a deeper understanding of the details of his data. (It appears that the interactive graphic computer may provide a useful tool that will allow the analyst to "zoom" from the details of his data up to a summary provided by a small number of cluster centers and the characteristics of the clusters.)

(d) Clustering can be used as a data manipulation algorithm. It can be used to reduce the dimensionality by first finding k cluster in a data space of p greater than k dimensions and from this obtaining the k-1 dimensional subspace of the original data "spanned" by these cluster centers. Whether other clustering algorithms provide a useful alternative to more traditional methods of principal components or factor analysis has yet to be shown.

(e) Clustering can be used to reduce the volume of a large data set by substituting for, say, 5,000 data samples, a non-random representative set of, say, 150 cluster centers that will tend to relatively adequately characterize the variation of the data. This similar set of cluster centers can then be used as input to such techniques as the multidimensional scaling programs that are not able to handle such large data volumes.

(f) In its loose role in exploring data, clustering can aid concept formation by providing a tractable description that still retains some of the subtleties of the data that are lost when the data is described by a single mean and covariance matrix or other such simple descriptors. These techniques are useful for exploring the data and for examining relationships between the variables and how the particular data population chosen affects relationships between the variables. For example, the existence of two distinct subgroupings may seem to imply a correlation between two variables in a single population, when in fact it is primarily indicating a correlational relationship between the means of two different subpopulations. In the same way the choice of variables affects the groupings found, and this realization forces the analyst to reconsider the measurements he has chosen and sometimes the weightings applied to those measurements in a way that may not be as easy to overlook as it is with some other methods.

Clearly it is the digital computer that has made the use of clustering techniques attractive and in many instances possible for a variety of people. Not only has it provided the computational power to actually perform the clustering algorithms but it has also put us in the position where we can collect a great deal of multivariate data. The existence of this data has led us to want to understand, to explore, to search for patterns of association in the data, and this in turn has fed back into our needs for some techniques which could organize this data. This underlines the statement that convenience is not only dramatically helpful but it is also dramatically demanding. Once we have been given the tool, we find uses to which it can be put.

3.1 Clustering as Fitting

Clustering can be viewed as a classical goodness of fit problem. For example, one can assume a model for the data of a mixture of multivariate distributions and then test the fit of this model to the data. This approach has not proved very tractable in terms of computing, but some recent work of Wolfe (1967) shows promise.

More generally, one assumes a certain kind of structure and then attempts to fit this structure to the data. The analyst then examines departures from the assumptions, that is, the residuals, and indicates changes that need to be made in the model. In some instances it may be possible to indicate the directions that changes need to be made in order for the model to fit more satisfactorily. For example, consider principal components. If one fits a p-dimensional space with k less than p subspace, then the residuals for the individual data samples are well defined in that they are the distance between the data point in the subspace and the data point in the entire space.

In many types of clustering techniques it is not clear what the appropriate notion of goodness of fit is, and the notion of the residual needs to be further developed.

3.2 Evolution of Clustering Procedures and Criteria

Initially there was the search for algorithms that in some intuitively satisfying way found groups in the data. This work on formal clustering algorithms started as early as Tryon (1939) and is still going on, as exemplified by the ISODATA algorithm of Ball and Hall (1967), by the algorithms of Sokal and Sneath (1963), and by the k-means algorithms of MacQueen (1967). As these algorithms were used, there developed a concern as to the criteria being used and attempts to optimize criteria, as in the work of Ward (1963) and Singleton. The work of Hartigan (1967) is a further extension in which he formulates criteria for hierarchical structure. Then the focus shifted toward the development of algorithms to optimize the criteria found to be useful. This is exemplified by the $|T|/|W|$ criterion used by Friedman and Rubin (1967). The acceptance of this criterion shifted emphasis toward developing algorithms that would optimize that criterion.

Thus we see the need for clustering techniques expressed initially in a search for algorithms that would produce "clusters". By and large at this stage it was considered that exhaustive computation of all potentially useful partitions of the data was impossible, and the goal was to get one best answer as defined by the algorithm. Gradually, as awareness increased through the use of the algorithms, it became possible to see that criteria were in fact being optimized in many cases [see J. Gower (1967), where he shows that different algorithms lead to the same underlying structure]. As computational power increased sufficiently, it was, at times, possible to evaluate exhaustively the criteria, usually in the case of clustering for small problems. Finally the work then moved into developing algorithms dominated by a criterion where the focus now is on non-exhaustive search procedures that tend to optimize the criterion [see Friedman and Rubin (1967)].

It is clear that there is a large variety of techniques if one only looks at the algorithms used to cluster the data. But it is not equally clear that the groupings that result from using these algorithms will differ substantially. The implicit criterion controlling how groups are found that is defined by the very nature of the algorithm or by the measure of similarity that is used may not be substantially different from algorithm to algorithm. For example, the ISODATA algorithm (1967) and the Singleton-Kautz algorithm (Singleton, 1967) appear to be very different. And yet when, upon finding the groupings obtained to be very similar, we examined the algorithms further, we found that the Singleton-Kautz algorithm was explicitly attempting to find a minimum variance partition while the ISODATA algorithm was implicitly tending to find that same type of partitioning.

In Friedman and Rubin (1967) we see the beginnings of the comparison of different criteria against the same data sets. In this paper the relation of some of the clustering methods with traditional multivariate statistical theory is elucidated. In this context a paper by Peter Ihm (1965) is very relevant.

In the area of factor analysis, a recent paper by M. Browne (1968) studies statistical properties of several methods of obtaining estimates from data. Again this paper is indicative of what is happening in cluster analysis. Different procedures were developed, and some amount of synthesis is take place, with people beginning to compare and evaluate different methods on data with known structures.

A striking example of the interaction of the methods of principal component analysis, factor analysis, cluster analysis and discriminant analysis applied to the analysis of psychiatric data is given by Friedman and Rubin (1968).

In the future, as computing power grows and changes in character with the development of the interactive graphic computer, it will be interesting to see if, in fact, algorithms are developed that are suitable for control by a man in a man-machine environment. In a much broader context, as people begin to develop on-line computer systems for data analysis, it is becoming increasingly apparent that one or more methods of cluster analysis will be included as data analysis tasks. It should be emphasized that in analyzing most large sets of multivariate data different methods of analysis will be employed on the same data set, and some form of cluster analysis will be most helpful in assessing the nature of the heterogeneity of the data as well as to group the data into more homogeneous groups for analysis. Further, as shown in Friedman and Rubin (1968), once having found some group structure, further multivariate analysis may be required to describe the data in a way that can be meaningfully interpreted in the paradigm of the subject matter of the data. The problem of interpreting the results of clustering procedures is very much a problem that still exists in the more traditional methods of multivariate analysis. We have a long way to go in our understanding of multivariate data.

4.0 Some Examples

In this section we consider three separate sets of data, each of which illustrates a particular aspect of using clustering techniques on real experimental data.

4.1 The Indian Studies Data

The desired goal of this clustering that used the Singleton-Kautz algorithm was to obtain a representative (non-random) sample of Indian cities that would allow a small number, say 10%, of the cities for detailed study as to the economic effects of improving their infrastructure, that is, their sewage system, electric power network and road structures. The aspect of this clustering that we would like to emphasize is shown in Figure 4, in which we have displayed the contribution of the sum of squared error of each of nine variables used in this particular clustering. (We clustered using a variety of sets of clustering variables. Figure 4 shows the result of one particular clustering.) The curves shown are essentially a "decomposition" of the total sum of squared error curve. For example, we see that in increasing the number of clusters from one cluster to two cluster that the contribution of Variables 1, 2, and 3 to the total sum of squared error markedly decreased with most other variables not being substantially decreased in their sum of squared error. In increasing from two to three clusters we see that Variables 6 and 9 cooperated, and in going from three to four clusters we see that Variable 5 contributed markedly, with Variable 1 contributing somewhat. This ability to perceive the way in which variables interact in reducing the sum of the squared errors suggests that strong relationships do exist between these variables. With this information, it is then possible to go back and examine the details of that interrelationship for just those three variables, using clustering or perhaps even graphic techniques. (This desire to go back and forth has been a major factor in the development of PROMENADE--a multivariate data analysis system that uses interactive computer graphics. See Ball and Hall, 1967.)

These relationships between variables and the observation that pairs or triples of variables tended to cooperate suggested to us that we ought to re-examine the relationship between principal components analysis and minimum variance clustering partitions. This comparison is still in process and has not yet been completed.

4.2 Job Satisfaction Data

This data consisted of 209 responses by Air Force scientists to questions that sought to ascertain the degree to which they were satisfied with their present employment and to elicit some of the factors related to satisfaction. One useful grouping consisted of seven clusters. We now describe the average response patterns for two of these clusters in order to show the suggestiveness of the groupings in suggesting explanations for job satisfaction. The work profile for Group 7 describes the people of this group as being between 35 and 50 years of age (95%); highly educated (50% Ph.D., 50% M.S.); civilians primarily performing basic research (80%); planning to work for the employing organization in ten years; publishing professional papers frequently (75% have five or more in the last five years); quite satisfied with their jobs (55% highly satisfied and only 15% dissatisfied). This can be compared to the work profile of Group 6, where 20% are under 30 years of age, have a moderate degree of education (33% M.S., 67% B.S.); are civilians; primarily performing applied or basic applied research (60%); planning to be working in the same organization in ten years; publishing infrequently (45% have zero to two publications in the last five years); only moderately satisfied with their jobs (30% highly satisfied and 35% dissatisfied). Another grouping found was that of the young researcher who had recently obtained his

Ph.D. and was now fulfilling his military obligation working in an Air Force laboratory at relatively low pay compared to civilians working in the same capacity. This group tended to be highly dissatisfied.

This data set indicated that the high interrelationship between the various questions on this questionnaire did almost necessarily insure that a useful grouping would result in the clustering and that the profiles would allow for interpretation in a convenient fashion. This coherence of the variation measured by the questions on the questionnaire in this data set did not exist in the data set which we discuss next.

4.3 Controversial Issues Data

This data consisted of questions asked about attitudes on various controversial subjects, including questions about religion, sex, politics, and other personal attitudes. The questionnaire was potentially designed to cut across many different attitudes, and there was far less interaction between various questions on the data set.

We clustered the data using all of the questions and were unable to discover any useful or interesting clusterings. We then focused on one set of questions relating to attitudes about sex. Out of this focusing on about one quarter of the questions, we were able to obtain extremely well-defined clustering that had a simple and straightforward explanation.

This data set led us to conclude that there is a great need for methods that search for variables on which to do the clustering in addition to the already existing techniques such as principal components. In a broader sense, the results suggest that there is a kind of interaction between the variables used and the objects used that suggests iterative procedures that successively define subpopulations both of objects and variables that lend themselves to interpretation. The task would then become one of interrelating results obtained on the different subpopulations--perhaps by using the subcategorizations obtained from considering just subsets of the variables and and subpopulations of objects.

5.0 Summary Remarks

Computers have added conceptual as well as algorithmic dimensions to statistics, and clustering techniques are one evidence of this change.

The movement caused in statistics by the use of the computer has led to new flexibility in criteria and the acceptance of more empirical methods. For example, other criteria than least squares are now acceptable. This movement makes it easier to accept clustering techniques as part of the statistics tradition.

Clustering itself is in transition. Movement is taking place from a loosely structured body of ad hoc algorithms toward a coherent set of tools whose interrelationships with each other and with other existing multivariate techniques are becoming known.

References on Clustering

1. Ball, Geoffrey H. (1965), "Data Analysis in the Social Sciences", American Federation of Information Processing Societies Conference Proceedings, Fall Joint Computer Conference, Vol. 27, Part 1, Washington: Sparton Books, London: Macmillan, pp. 533-560.

2. Ball, Geoffrey H. and David J. Hall (1967), "A Clustering Technique for Summarizing Multivariate Data", Behavioral Sciences, 12, No. 2, pp. 153-155.

3. Ball, Geoffrey H. and David J. Hall (1967), "PROMENADE, An On-Line Pattern Recognition System", Stanford Research Institute Technical Report No. RADC-TR-67-310, Menlo Park, California, pp. 1-126.

4. Berry, Brian J.L. (1960), "An Inductive Approach to the Regionalization of Economic Development", #62, University of Chicago, pp. 78-107.

5. Bonner, R.E. (1964), "On Some Clustering Techniques", IBM Journal, 22, pp. 22-32.

6. Browne, Michael (1968), "A Comparison of Factor Analytic Techniques", Psychometrika, Vol. 33, No. 3.

7. Dale, A.G. and N. Dale (1965), "Some Clumping Experiments for Associative Document Retrieval", Am. Documentation, Vol. 16, No. 1, pp. 5-9.

8. Dalenius, Tore and M. Burney (1951), "The Problem of Optimum Stratification II", Skandinavisk Aktuarietidskrift, Vol. 34, pp. 133-148.

9. Eusebio, James W. and Geoffrey H. Ball (1968), "ISODATA-LINES--A Program for Describing Multivariate Data by Piecewise-Linear Curves", Proceedings of International Conference on Systems Science and Cybernetics, University of Hawaii, Honolulu, Hawaii, pp. 560-563.

10. Fisher, Walter D. (1965), "Constructive Partition and Aggregation", Kansas State University, Manhattan, Kansas, pp. 1-19.

11. Forgy, Edward W. (1963), "Detecting 'Natural' Clusters of Individuals", Department of Psychiatry, University of California Medical Center, Los Angeles, pp. 1-10.

12. Fralick, S.C. (1967), "Learning to Recognize Patterns without a Teacher", IEEE Transactions on Information Theory, IT-13, No. 1, pp. 57-64.

13. Friedman, H.P. and J. Rubin (1967), "On Some Invariant Criteria for Grouping Data", Journal of the American Statistical Association, 62, No. 320, pp. 1159-1178.

14. Friedman, H.P. and J. Rubin (1968), Chapter 5 "Logic of Statistical Procedures", The

Borderline Syndrome, by Grinker et al., Basic Books.

15. Gower, J.C. (1967), "A Comparison of Some Methods of Cluster Analysis", Biometrics.

16. Green, P. et al. (1968), "Analysis of Marketing Behavior Using Nonmetric Scaling and Related Techniques", Technical Report, Marketing Science Institute, University of Pennsylvania.

17. Guttman, Louis (to be published December 1968), "A General Non-Metric Technique for Finding the Smallest Coordinate Space for a Configuration of Points", Psychometrika.

18. Hartigan, J.A. (1967), "Representation of Similarity Matrices by Trees", Journal of the American Statistical Association, Vol. 62.

19. Ihm, P. (1965), "Automatic Classification in Anthropology", The Use of Computers in Anthropology, Edited by Dell Hymes, London: Mouton and Company.

20. Johnson, Stephan C. (1967), "Hierarchical Clustering Schemes", Psychometrika, 32, No. 3, pp. 241-254.

21. Kruskal, J. (1964), "Multidimensional Scaling by Optimizing Goodness of Fit to a Non-Metric Hypothesis", Psychometrika, 29, pp. 1-28.

22. Kruskal, J.B. (1964), "Nonmetric Multidimensional Scaling: A Numerical Method", Psychometrika, 29, No. 2, pp. 115-129.

23. MacQueen, J.(1967),"Some Methods for Classification and Analysis of Multivariate Observations", 5th Berkeley Symposium on Mathematics, Statistics and Probability, 1. pp. 281-297.

24. Mattson, R.L. and J.E. Damman (1965), "A Technique for Determining and Coding Subclasses in Pattern Recognition Problems", IBM Journal, pp. 294-302.

25. Medgyessy, Pal (1961), Decomposition of Superpositions of Distribution Functions, Publishing House of Hungarian Academy of Science, Budapest.

26. Nunnally, Jim (1962), "The Analysis of Profile Data", Psychological Bulletin, 59, No. 4, pp. 311-319.

27. Rao, C.R. (1964), "The Use and Interpretation of Principal Component Analysis in Applied Research, Sankhya, Series A, Vol. 26, Part 4, pp. 329-358.

28. Rogers, David J. and Taffee T. Tanimoto (1960), "A Computer Program for Classifying Plants", Science, 132, pp. 1115-1118.

29. Rubin, J. (1967), "Optimal Classification into Groups: An Approach for Solving the Taxonomy Problem", Journal of Theoretical Biology, 15, pp. 103-144.

30. Sammon, John W.,Jr. (1968), "An Adaptive Technique for Multiple Signal Detection and Identification", Pattern Recognition, Edited by L. Kanal, Thompson Book Company, Washington, D.C., pp. 409-439.

31. Sebestyen, George S. (1966), "Automatic Off-Line Multivariate Data Analysis", Proceedings of Fall Joint Computer Conference, Spartan Books, pp. 685-694.

32. Shepard, R.N. (1962), "The Analysis of Proximities: Multidimensional Scaling with an Unknown Distance Function I, Psychometrika, 27, pp. 125-140.

33. Singleton, R.C. (1967), Private Communications, Stanford Research Institute, Menlo Park, Calif.

34. Sokal, R.R. and P.H.A. Sneath (1963), Principles of Numerical Taxonomy, W.H. Freeman and Company, San Francisco.

35. Stanat, Donald F. (1968), "Unsupervised Learning of Mixtures of Probability Functions", Pattern Recognition, Edited by L. Kanal, Thompson Book Company, Washington, D.C., pp. 357-389.

36. Stark, Lawrence, Mitsuharu Okajima and Gerald M. Whipple (1962), "Computer Pattern Recognition Techniques: Electrocardiographic Diagnosis", Communications of the Association for Computer Machinery, 6, No. 10, pp. 527-532.

37. Torgerson, W. (1968), "Multidimensional Representation of Similarity Structures", The Role and Methodology of Classification in Psychiatry and Psychopathology, Edited by Katz, Cole and Barten, U.S. Department of Health, Education and Welfare.

38. Tryon, Robert C. (1967), "Person Clusters on Intellectual Abilities and on MMPI Attributes", Multivariate Behavioral Research, Vol. 2, pp. 5-34.

39. Tucker, Ledyard R. (1968), "Cluster Analysis and the Search for Structure Underlying Individual Differences in Psychological Phenomena", Conference on Cluster Analysis of Multivariate Data, New Orleans, Louisiana, pp. 10.01-10.17.

40. Wallace, David (1968), "Cluster Analysis", International Encyclopedia of the Social Sciences, Cromwell Collier, pp. 519-524.

41. Ward, Joe H., Jr. and Marion E. Hook (1963), Application of an Hierarchical Grouping Procedure to a Problem of Grouping Profiles", Educational and Psychological Measurement, 23, No. 1.

42. Wolfe, John H. (1967), "NORMIX--Computational Methods for Estimating the Parameters of Multivariate Normal Mixtures of Distributions", Technical Report, U.S. Naval Personnel Research Activity, San Diego, California, pp. 1-31.

STATISTICAL TECHNIQUES IN CLUSTERING AND PATTERN RECOGNITION

Paul Switzer
Stanford University

1. Of the many different problems in clustering and pattern recognition we have chosen to examine several which have the following common features:

> The objective is to partition the individuals of a specified population into subsets.
>
> The permitted partitions may depend only on the values of a specified measurement variable X defined for this population.

2. The properties above imply that our problems are equivalent to partitioning the space of all possible values of X; in particular, two individuals with the same X value must not be in different subsets of the partition. Quite a variety of clustering and pattern recognition problems can be regarded in this framework and they are distinguished by what we hope to accomplish with our chosen partition.

3. Definitions: A subset of X values will be called a stratum, especially when it refers to one of the subsets of our chosen partition. A collection of individuals whose X values all belong to the same stratum will be called a cluster. So, for a given stratification of X and a given group of individuals, there is a unique decomposition of the group into clusters. The definition of a cluster and a stratum are equivalent only in the case of a finite population where no two individuals assume the same X value.

4. Often, what we hope to accomplish with a chosen partition is not well articulated, as we shall see later. Occasionally, we can be quite explicit however, as the next several examples will illustrate.

5. Equal probability partitions. Here the objective is to get k strata each with probability content $1/k$; i.e., each stratum contains an equal fraction of the population. Without further requirements this objective is easily achieved, in general, if we know enough about the distribution of X values. The problem becomes interesting when this distribution is unspecified. Suppose X has a continuous distribution on a Euclidean space, and we have a random sample of m individuals with observed values $X_1, X_2, \ldots, X_m$; Fraser [1] has given a general class of partitioning procedures, based on the sample, with the property that the expected probability content of each of the strata is $r/(m+1)$ where r is any specified divisor of m+1. If X is real-valued, for example, the sample points themselves partition the line into m+1 intervals, each with expected probability content $1/(m+1)$.

6. As a second example with an explicit objective consider the k-means problem - as it is sometimes called. What we need to specify is k, the number of strata, and a distance function ρ defined on the space of X. Once ρ has been specified, the mean μ_h of stratum h is defined to be that point in X space which minimizes $V_h = E[\rho^2(X,\mu_h) | X \in \text{stratum } h]$. If W_h is the probability content of stratum h, then the total dispersion within strata is taken to be $V = \Sigma W_h V_h$. The objective is to choose a partition which minimizes this total within strata dispersion. (For example if ρ is Euclidean distance then μ_h is the usual stratum center of gravity.)

7. MacQueen [2] has discussed this problem and indicated how it might arise in real situations. One illustration not developed there concerns estimation of the mean of X when X is real-valued: If the allocation of a random sample to strata is proportional to the stratum weights, then the mean-square-error of the sample mean is indeed proportional to V if $\rho^2(a,b)$ is taken to be $(a-b)^2$.

7a. In practice, when we are estimating the unknown mean value of X in a population, it will not be possible to stratify the sample on the values of X itself. Instead what is usually done is to find a variable Y which is strongly correlated with X and whose distribution is pretty well known, then stratify according to Y values. In principle, there exists an optimum partition of Y which will minimize the V quantity for the X variable.

8. If the allocation of the sample to strata in the preceding example is according to Neyman optimal allocation (see Cochran [3]) then the appropriate quantity to minimize is $\Sigma W_h \sqrt{V_h}$. Dalenius and Hodges [4] have, for this case, obtained an approximate solution to the partitioning problem: On the assumption that k is fairly large, the distribution of X (assumed continuous) will be nearly uniform within strata. Using this uniformity, the optimum stratum boundaries can be approximated by taking equally spaced points on the scale

$$Z(X) = \int_{-\infty}^{X} \sqrt{f(t)}\,dt ,$$

where f is the density function of X.

9. We now turn to some partitioning problems where the objective is not so explicit; the ones we will be discussing frequently go under the name of cluster analysis. When one tries to

articulate his objective in doing cluster analysis, one tends to come up often with something like - to chop up a supposed multimodal distribution of X values by putting one mode in each stratum. Often underlying this objective is the notion of trying to represent the population as a mixture of intrinsically interesting subpopulations - each with its distinctive distribution of X values.

10. We might try to model the situation of the last paragraph in the following approximate way: There are k simply connected sets (islands) in X-space which are at a positive distance from each other and on which the density of X is positive; everywhere else the density is zero. See Fig. 1. The objective is to have each stratum of our chosen partition contain one island. It is assumed, of course, that a distance has been defined on X space. If the population is finite this model may not seem too helpful, but in such cases it is sometimes possible to think of the population as having been sampled from a superpopulation with a continuous X distribution of the above type. This model is surely an over-simplification of most real situations; but if our partitioning procedures can not do well in such simplified problems, they are not likely to do well in the less well resolved practical situations. Such a model might, therefore, be useful in a comparative study of partitioning procedures.

11. If we had this kind of population in mind and a random sample from it, what sort of procedures are reasonable for dividing the sample into clusters (recall the definitions of Par. 3). If the sample is correctly clustered, then we might expect the individuals within a cluster to have X values close to each other, whereas distances between clusters should be fairly large. Considerations of this kind have led various people to fix on within-or between-cluster distance criteria as a means of evaluating any given partition. We will call such criteria C criteria, generically, and give some examples

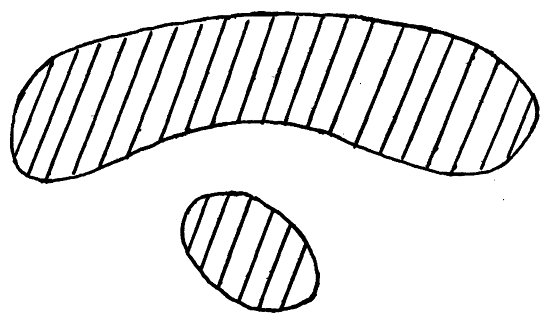


Fig. 1. The shaded region is where the density of X-values is positive; it is an example of the model mentioned in Par. 10.

below. The corresponding procedure then tries to find a partition which optimizes (maximize or minimize) the criterion C - as a substitute for the less explicit objectives of Par. 9 and 10.

12. One example of a C criterion is the total within strata dispersion, the quantity V of Par. 6. A similar more widely used C criterion is W, the determinant of the <u>pooled</u> within strata covariance matrix, for X variables defined on a Euclidean space. Specifically, if S_h is the covariance matrix conditional on X being in stratum h, and W_h is the weight of stratum h, then $W = \det(\Sigma W_h S_h)$. We would then try to find a minimum W or a minimum V partition. Minimizing W or V also has a tendency to maximize the average distance between strata as measured by the distance between the stratum means. The V and W and similar criteria have been explored and used in this way by Friedman and Rubin [5] among others. For example, the criterion trace $(\Sigma W_h S_h)$ is sometimes suggested; it is not often recognized that this is equivalent to MacQueen's V when Euclidean distance is used to compute V.

13. Whether minimizing V or W gives us the kind of partition we really want is at least sometimes open to question. Unfortunately, it is all too easy to construct examples where such minimizing does not divide up multimodal distributions in the desired way. In particular, let X be real valued and have zero density everywhere except on the intervals $(0,.3)$ and $(.4, 1)$ where it is constant. See Fig. 2. If we were smart enough to have chosen $k = 2$, then the articulation of the objective of Par. 9 would be to partition the line into two half-lines with the boundary point between 0.3 and 0.4. However, this partition actually gives a larger (worse) value for the V or W criterion than the partition with boundary point at 0.5, say. One might well regard this as peculiar or unsatisfactory.

14. Another example is illustrated in Figure 3. It shows a finite population of 10 individuals taking on X values in the unit square. Two selected partitions for k=2 are also shown. Although we might feel happier with the first of the two partitions, it is the second one which gives a better (smaller) value for V or W. This example hints at the care one must exercise in choosing a C-criterion, if this is the desired approach. Part of the problem here is that both V and W may not work well when the islands differ appreciably in size or shape.

15. The actual calculation of an optimum C partition, given the distribution of X, is in general a very difficult mathematical task. If the population is finite, or if it is infinite but there are only a finite number of different X-values in the population, then the number of possible partitions of the X values into k sets is also finite. In principle, it is then possible to compute C for all possible partitions and thus find the optimum C partitions.

However, if there are m different X-values then there are $k^m/k!$ partitions to enumerate for each fixed k - an unreasonable task if m is any size at all.

16. Instead of enumerating to get an optimum C partition in this finite case, Fortier and Solomon [6] have examined the possibility of using the partition which optimizes C among a random sample of partitions. Suppose there are about 10^{10} possible partitions and we take a sample of 10^3 of them. The probability that the sample will contain the overall optimum C partition (assume it is unique) is still minute, viz. 10^{-7}; but the probability that the best partition in the sample is among the best 1% of all partitions is very high, viz. $1-(.99)^{1000} = .99995$.

17. Despite the fact that the sampling method is likely to give a partition with a very good C, it seems from the Fortier and Solomon study that very good is not good enough. In the first place partitions with even better C can be obtained by various simple ad hoc procedures to be discussed in Par. 32. In the second place it seems that the distribution of C values has a very long thin tail, or that even slightly sub-optimal C values may correspond to unappealing partitions.

18. Although Fortier and Solomon used a particular C criterion not discussed here, we feel that these conclusions about the sampling of partitions have more general validity. However, if we know that an optimum C partition must have certain properties, then we can restrict our sampling to those partitions possessing such properties. For example, the clusters of an optimum partition may have disjoint convex hulls. There will be relatively very few such partitions and our sampling effort is likely to be much more successful when so restricted. Of course, it is probably hard to come by such properties of optimal C partitions; but we can make them up anyway so as to be consistent with the real objective (See Par. 9) - which, after all, is <u>not</u> the optimization of C.

19. As a simple example consider the finite population of Figure 3 again. With k=2 there are 512 possible unrestricted partitions. With the convexity restriction of the preceding paragraph there remain only m(m-1) partitions to consider, or in this example only 90.

20. Other approaches to find optimum C partitions have been explored by Friedman and Rubin [5] and a good list of references is contained there. Common to these approaches is that one starts with a given partition, modifies it according to a fixed procedure, and ends up with new partition with a better C value. The modification procedure is iterated until it reaches a stable partition, i.e., one which is unaffected by the modification. It is then said to have converged and the resulting stable partition becomes a candidate. For example, for a given partition into k clusters of a finite population in Euclidean space, the cluster means and the pooled within clusters covariance S are computed. The Mahalanobis distance of each individual to each of the cluster means is then also computed. The modification procedure consists of reassigning each individual to the cluster to whose mean it is closest. This guarantees that the value of the criterion function W of Par. 12, the determinant of the pooled within clusters covariance matrix, will be reduced. However, there are many stable partitions for any given configuration of X values, some good and some not so good. Any one of them can be reached by a suitable input partition. Of course, the minimum W partition is a stable partition, also.

21. Although the modification procedure guarantees a sequence of partitions with decreasing W, the final stable partition of the sequence may be nowhere near a minimum W partition. Of course, the modification procedure could be used with several different input partitions, then we can choose the best of the output stable partitions. Indeed, such modifications routines can be combined with Fortier-Solomon sampling of input partitions. Figure 3 shows the final result of using the modification procedure of the last paragraph, starting from two different initial partitions. A recent application is reported by Demirmen [7] who used the technique to "improve" rock classifications. Demirmen also wrote an efficient IBM 360 computer program to implement the technique.

22. In summary, we see that replacing a general clustering objective of the type in Par. 9

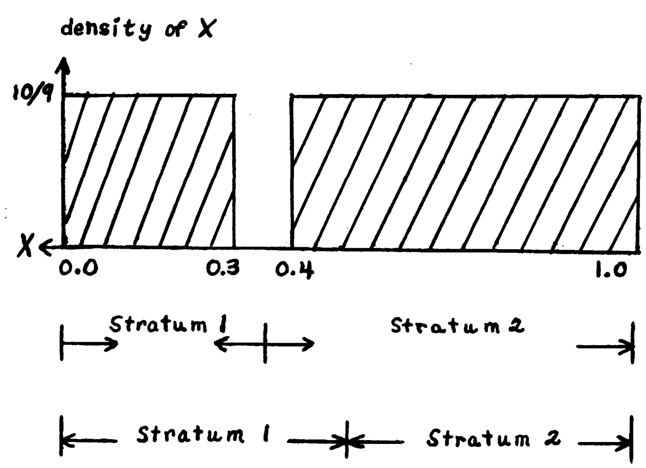


Fig. 2. The X distribution of Par. 13 and two different partitions of X. If we were trying to minimize W (or V), the second partition would be judged better.

by a specific objective which seeks to optimize the value of V or W may or may not be a reasonable thing to do. The state of the art is such that we don't always know when it is reasonable. Furthermore, we don"t yet have proven efficient methods for finding optimum V or W partitions, even when we do think it is reasonable.

23. As an alternative to V or W, we might look for a C criterion which is especially tailored to do well for models of the type in Par. 10. Recall that the islands are assumed to be a positive distance apart. It seems more reasonable then to measure distance between strata in terms of the minimum distance between any pair of individuals from different strata - rather than in terms of the distance between stratum means which is implied in the use of V or W.

24. Specifically, let the minimum distance between stratum h and stratum i be denoted by R_{hi}. For any subgroup of $m > 1$ individuals there will be a subdivision into two strata which is optimal in the sense that R is maximized over all partitions of the m X values into two sets. This maximum value of R could be taken as a measure of the within subgroup distance for that subgroup of m individuals. For example, in Figure 4 the within distance for cluster 1 is 21mm, for cluster 2 it is 26 mm, and the R distance between the two clusters is 22mm. (Using Euclidean distance.) For a chosen partition of the whole population we define R_h to be the within stratum distance for stratum h. If a stratum has only one point take $R_h = 0$. As an alternative to W type criteria, we suggest using something like $Z = \min R_{hi} - \max R_h$. The objective now is to find the maximum Z partition of the population.

25. Maximizing criteria like Z is likely to be consistent with the clustering objectives of paragraph 9 and 10; especially for fairly large samples. This is suggested by the following consideration: Suppose our partition did correspond to the "islands" of the model, then splitting an island would greatly decrease $\min R_{hi}$ whereas merging two islands would greatly increase $\max R_h$. The merit of a Z criterion should not be much affected by diversity in the sizes and shapes of the islands or the shapes of the distributions within islands. The maximum Z partition for the population of Figure 3 is the first one shown there (A).

26. Z also has the satisfying property that, for any partition, it is computible from the pairwise distances between the points of the sample. In fact it has the property that for any specified number k of strata, there is a unique partition for which the Z value is positive. The last statement assumes there are no ties among the pairwise distances and implies that this unique partition is the optimum one. The proof is a little involved and rests on the following observations: If any sample is divided into two clusters so that the R distance between them is maximized then this is an optimum Z partition for k=2, and Z is positive; the optimum Z partition for $k' > k$ strata is always a refinement of the optimum Z partition for k strata; if we have subsets of two strata then the distance between the subsets is not less than the distance between the strata; finally if a single stratum is divided into several subsets, then there will be a pair of subsets with distance betweeen them not less than the within stratum R-distance as defined in Par. 24. The construction of a proof is now left to the reader.

27. The refinement property of maximum Z partitions, noted in the last paragraph, suggests a simple algorithm for finding them. For, suppose we have the best partition for k+1 strata; if we merge the two nearest strata (in terms of R distance), it follows that we will now have the best partition for k strata. In particular, we can start by assuming that each of m distinct X-values of the sample is a cluster unto itself. Then, by repeating the process of merging nearest clusters m-k times, we will end up with the maximum Z partition for k clusters. So we see that this particular criterion can be optimized in an extraordinarily simple manner. It also follows that the quantity $\max R_h$ for optimum k clusters is equal to the quantity $\min R_{hi}$ for optimum k+1 clusters, hence the sequence of maximum Z values for each k is itself easily computed. This would be useful if our choice for the number of strata was based on comparing Z values for different k.

28. There is an even more direct and general approach to the partitioning problem when the population is believed to resemble the model of Par. 10. Suppose we could estimate the density of X somehow. Then the estimated "islands" could be those connected sets in X-space on which the estimated density exceeded a certain threshhold value; the chosen partition would then put one island in each stratum. Note that the number of strata is not fixed in advance. A naive method of threshholding the density at a point X_o in X space is to count the number of sample points in a fixed neighborhood around X_o. This count can then serve as the threshholding device. This procedure has been applied to the finite population of Figure 3 with circular neighborhoods of radius 11 min and a threshhold of 1 point. The resulting partition is shown in Figure 5.

29. The density threshhold approach and the maximum Z approach are both clearly tailored to the partitioning of populations which are more or less of the type in Par. 10. We have tried to argue that clustering of populations is often attempted only because it is believed that such a description roughly obtains. But it is clear that both of the last named approaches can fall down in many situations.

30. For one, if the sample is small both procedures become rather sensitive to the actual configuration of sample X values. The maximum Z approach is, in general, sensitive to outlying X values; this could possibly be remedied by weighting the R distances by the number of points involved in the pair of clusters, or a similar device. A problem with the model occurs when there are relatively high density "bridges"

between pairs of islands. See Figure 6. We may still want such a pair of islands to show up in different strata, but the tendency of our last suggested procedures would be to lump such connected islands into a single stratum, even in large samples.

31. We now return to re-examine the algorithm used to find maximum Z partitions, as described in Par. 27. This algorithm belongs to an interesting class of stepwise partitioning algorithms very much in the spirit of the one used by King [8] and by Sokol and Sneath [10]. We will call algorithms of this type stratum-merging procedures.

32. Suppose we already have a partition of the population into k strata (based on X values), then we can obtain a partition into k-1 strata by merging two of the existing strata into a single stratum. If we specify a function D which measures distance between any pair of strata, then the natural thing to do would be to merge the two strata which are nearest each other. Such a merging procedure can be iterated to obtain a partition with any number of strata less than k. In particular, for populations taking on only a finite number m of different X values, we can get started by considering each X-value to be a stratum by itself. If we iterate the merging routine with a specified D function

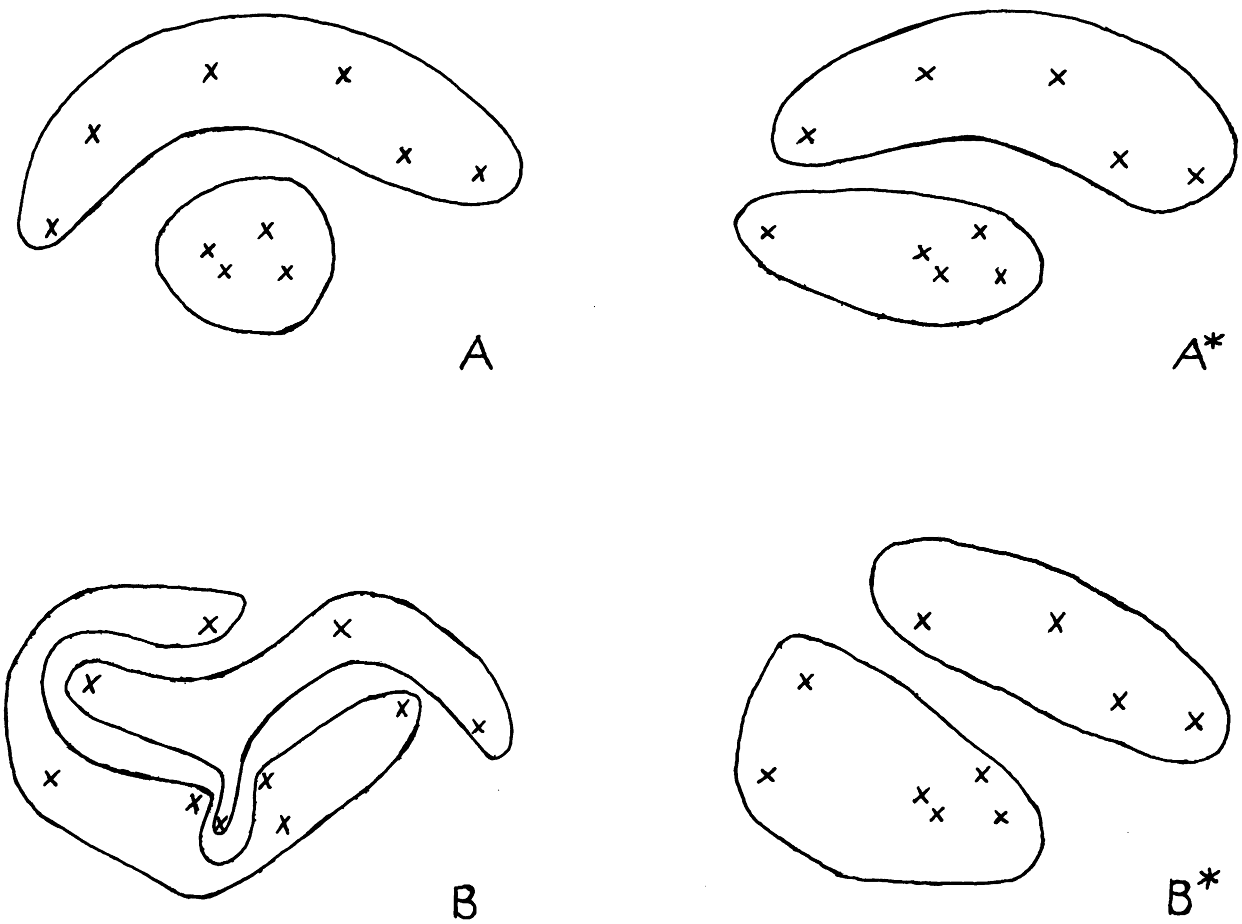


Fig. 3. Four partitions of a population of ten individuals into two clusters. The partitions A* and B* are stable by the method of Par. 20 and can be reached from initial partitions A and B, respectively, in one iteration. The W values are (in increasing order) .0047 - B*, .0056 - A*, .0158 - A, .0388 - B. The V values are (in increasing order) .2515 - B*, .3341 - A*, .4112 - A, .4662 - B. The only partition with a positive Z value is partition A, with Z = .04. See Par. 24. The scale is 1 inch = .24 units.

exactly m-k times, we will then end up with exactly k strata - for any specified k.

33. It is clear that when D is taken to be the R-distance of Par. 24, then this exactly describes the procedure used to find maximum Z partitions. But for any arbitrary D measure it may not be evident which, if any, criterion is being maximized by such a stepwise procedure (in the sense of Par. 11). Indeed, the V or W criterion discussed in Par. 12 are unlikely to be optimized in this way for any choice of the cluster distance function D: We might try anyways by choosing D to be some measure of distance between the stratum means - but the stepwise procedure will tend to maximize the <u>minimum</u> distance between any pair of means, whereas V or W tends to maximize the <u>average</u> distance between means. Notwithstanding these divergent tendencies, the stepwise procedure may still give as good a value of V or W as that obtained by the more cumbersome methods of Par. 16 or Par. 20. Also, such stratum merging procedures are always relatively quick and simple to execute, and have the virtue of automatically generating a whole sequence of suggested partitions, one for each possible value of k, the number of strata.

34. Solomon [9] has obtained satisfactory partitions using cluster-merging on at least three sets of data, in each case basically using distances between means to measure distances between strata. King himself was partitioning real-valued variables $(Y_1, Y_2, \ldots$ say), and for a distance measure between clusters A and B he used

$$D^2 = \Big(\sum_{j\epsilon B}\sum_{i\epsilon A} S_{ij}\Big)^2 \Big/ \sum_{j\epsilon A}\sum_{i\epsilon A} S_{ij} \sum_{j\epsilon B}\sum_{i\epsilon B} S_{ij}$$

where S_{ij} is the covariance between Y_i and Y_j. This can be recognized as the square product-moment correlation between the average of variables in A and the average of variables in B. So, essentially, all he needs to be able to do is to estimate the covariance matrix.

35. Throughout the entire discussion thus far we have had freqeunt occasion to refer to the computation of distances - for the V criterion of Par. 6, for the definition of neighborhoods in the density estimation of Par. 28, and in many other places. Very little was said explicitly about how distances should be defined and it seems that little can be said. This is an especially thorny problem when the X variable is of high dimension and it becomes very easy to make very poor choices. We would probably want certain simple transformations of the X-variable to preserve the order relation among distances, and we might choose our distance function with this in mind. Monotonic transformations of the distance function will not affect minimum Z partitions; however, this is not true for the other procedures discussed here.

36. If the covariance matrix of a population in Euclidean space is S, then the distance between points X_o and X_1 in this space can be taken as $(X_o-X_1)'S^{-1}(X_o-X_1)$. This is what is called Mahalanobis distance with respect to the overall covariance; it is totally invariant under linear transformations of X. This distance has been used with satisfaction by Solomon [9] for example, but one can construct examples where it is obviously inappropriate. In general, we would have a much better idea of an appropriate distance measure if we already had the clusters, but this just begs the question. The cluster improvement procedures discussed in Par. 20 implicitly modify an arbitrary distance function to make it more appropriate.

37. The sensitivity of a partitioning procedure to the choice of a distance function will tend to decrease as the sample gets large and as the dimensionality of X goes down. While we may not be able to control our sample size, it might be useful to attempt to reduce dimensionality before fixing on a distance function. Broadly speaking we will need some luck in this anyways.

38. We have made no attempt so far to cast any of the foregoing partitioning problems in a decision theoretic framework. If we allow ourselves to postulate the existence of an "ideal" partition, then our loss could be measured by the extent to which the chosen partition differs from an ideal one. The measure of discrepancy should depend, of course, on the nature and objectives of the particular problem in hand. In general, it will not be possible to calculate the loss without knowing an ideal partition, and if we knew of such a partition we would naturally use it and incur a minimum loss. Losses are being introduced, therefore, not because they are a calculable quantity in real problems, but for two other reasons: first, to put the formulation of the objectives of partitioning procedures on a firmer conceptual basis; second, to provide a means of evaluating recommended partitioning procedures on artificial examples (with specified ideal partitions) as a clue to how they would perform on different kinds of real populations. Finally, in certain special cases involving random sampling from populations, it may actually be possible to calculate the <u>expected</u> loss of a given procedure, even when an ideal partition is not known.

39. As a first example consider the problem of partitioning a continuous X distribution into sets of equal probability content, the problem of Par. 5. If $P_1, P_2, \ldots, P_k$ are the actual (but unknown) probabilities of the k chosen strata, then our loss might be measured by $L_1 = \max P_h - \min P_h$. If we use the Fraser [1] method of partitioning into strata, it is actually possible to compute the expected loss for this method, since the P_h's so generated will have a completely specified Dirichlet joint distribution.

40. If the k-means problem of Par. 6 has the minimum mean-square-error objective of Par. 7, then it might be reasonable to measure the loss by something like $L_2 = \log[V_o/\min V]$ — where V_o

is the computable V value for the chosen partition and min V is generally not computable except in artificial examples. However, if we are trying to find minimum V partitions as an approximation to the general clustering objectives of Par. 9, it is probably inappropriate to measure the loss by L_2 or anything like it in which only V values are involved. A similar remark applies to procedures that look for partitions with small W values. See Par. 12.

41. Here is an approach to the specification of an appropriate loss function when general clustering objectives are involved. Let us call individuals who belong to the same stratum of an ideal partition of X - friends; individuals belonging to different strata of an ideal partition will be called enemies. If, in the chosen partition, an individual finds that he has f friends outside his stratum and e enemies within his stratum, his loss is e+f. The total loss, L_3, for the chosen partition may then be taken proportional to the sum of the e+f losses over all individuals in the population. As an example, if the partition A of Figure 3 is considered to be "ideal", then partition A* has loss $L_3/100 = .18$, B has loss .48, and B* has loss .32. For infinite populations the expected value of an L_3 type loss could be computed for a random sample of standard size.

42. The loss function of the last paragraph has an interesting characterization, whether for finite or infinite populations: Let $A_1, A_2, \ldots A_{k'}$ denote the strata of the ideal partition and let $B_1, B_2, \ldots, B_k$ denote the strata of the chosen partition. Note that k and k' are not necessarily the same. Now let P_i denote the probability content of A_i, Q_j the content of B_j, and P_{ij} the content of $A_i \cap B_j$. Then it turns out that L_3 is proportional to $\Sigma P_i^2 + \Sigma Q_j^2 - 2\Sigma\Sigma P_{ij}^2$. The proportionality factor would involve only the population or the sample size. Note that $L_3=0$ if and only if the chosen and ideal partitions coincide. As an example for an infinite population, suppose the ideal partition of the population in Figure 2 (k'=2) has its boundary point somewhere between X = .3 and X = .4; then the suboptimum partition with boundary at X = .5 (k=2) has $L_3 = 16/81$.

43. We conclude by extending the notion of an ideal partition. At the very outset we noted in Par. 1 that our chosen partition of a population must be based on the values of a specified measurement variable X. However, in populations where more than one individual can take on the same X value, it is a non-trivial extension to allow the ideal partition to be arbitrary, i.e., not to depend on X. The implication is that no partition based on X can be perfect in the sense that L_3 can never be zero, though there will still exist a best partition based on X with minimum L_3. The representation for L_3 of the preceding paragraph remains valid in this extended situation.

44. An example is provided if we want to cluster a human population into ethnic backgrounds but our measurement variable X is the surname. Note that this extension plays no role if we are strictly in the islands situation of Par. 10 and the objective is to isolate the islands. It is when the islands begin to overlap that the extension comes into play. One of the reasons for introducing the extension was to forge a link from the problems we have been discussing to the standard Wald assignment or classification problem. If the strata of the ideal partition have labels 1,2,...,k, and if k is specified, and if we are required to attach labels 1,2,...,k to the strata of our chosen partition — then we have arrived; i.e., it is possible to regard the difference between the Wald problem and the problems of this paper as the difference between having to choose an ordered versus an unordered partition of X space.

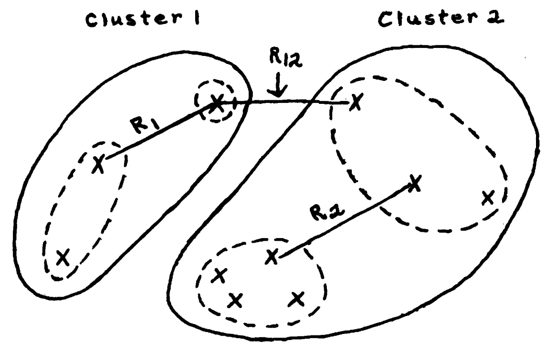


Fig. 4. This particular partition of the 10 individuals into two clusters has within cluster 1 dispersion R_1 = 21 mm, within cluster 2 dispersion R_2 = 26 mm, and between clusters dispersion R_{12} = 22 mm. This gives Z = -4 mm. See Par. 24.

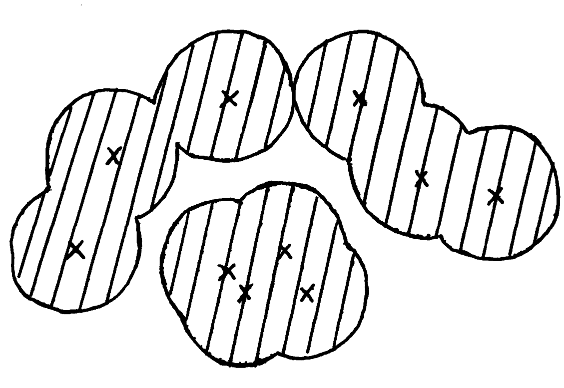


Fig. 5. The shaded area is where the density of X is estimated to be positive by the method of Par. 28 with circular neighborhoods of radius 11mm. The result is a partition into three strata as shown.

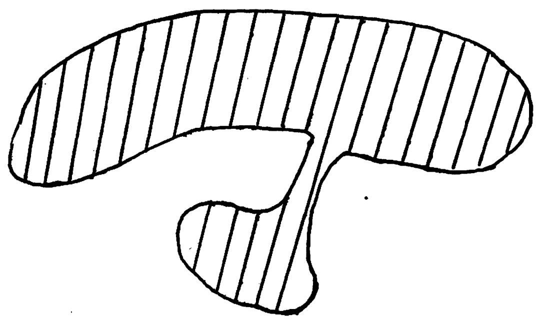


Fig. 6. The shaded region is where the density of X values is positive; it is an example of the model mentioned in Par. 30.

References

[1] Fraser, D.A.S., Nonparametric Methods in Statistics, John Wiley and Sons, Inc., New York, 1958.

[2] MacQueen, J., "Some methods for classification and analysis of multivariate observations," Fifth Berkeley Symposium (1965) 281-297.

[3] Cochran, W.G., Sampling Techniques (2nd edition), John Wiley and Sons, Inc., New York, 1963.

[4] Dalenius, T., and Hodges, J.L., Jr., "Minimum variance stratification," J. of Amer. Statist. Assoc., 54 (1959) 88-101.

[5] Friedman, H.P., and Rubin, J., "On some invariant criteria for grouping data," J. of Amer. Statist. Assoc., 62 (1967) 1159-1178.

[6] Fortier, J., and Solomon, H., "Clustering procedures," Multivariate Analysis, Edited by P.R. Krishnaiah, Academic Press, New York (1966).

[7] Demirmen, F., Unpublished Ph.D. dissertation, Geology Department, Stanford University (1968).

[8] King, B., "Stepwise clustering procedures," J. of Amer. Statist. Assoc., 62 (1967) 79-85.

[9] Solomon, H., personal communication.

[10] Sokol, Robert R., and Sneath, P.H.A., Principles of Numerical Taxonomy, W.H. Freeman and Company, San Francisco, 1963.

III

MEASUREMENT OF HOUSING QUALITY AND ITS POLICY IMPLICATIONS

Chairman, LAWRENCE N. BLOOMBERG, U. S. Office of Statistical Standards

METHODS OF MEASURING HOUSING QUALITY

Arthur F. Young and Joseph M. Selove, U.S. Bureau of the Census

I. Background

The term "housing quality" is, in a sense, a misnomer. The Census Bureau has never attempted to describe the complete spectrum of housing quality. Rather, our efforts have been directed to identifying that housing which, in the opinion of experts, constituted a serious hazard to health or safety. In the housing censuses of 1940, 1950, and 1960 indications of such conditions were provided in terms of the state of repair of the structure and the availability of plumbing facilities, that is, hot piped water in structure, flush toilet for exclusive use, and bathtub or shower for exclusive use. The state of repair of the structure was categorized in 1940 as "needing major repairs" or "not needing major repairs." In 1950, units were described as being "dilapidated" or "not dilapidated." In 1960, enumerators were instructed to determine whether a unit was "dilapidated" or "not dilapidated," and if "not dilapidated" then whether it was "deteriorating" or "sound." A dilapidated unit was described as one in which structural defects were either critical or so widespread that the structures should be extensively repaired, rebuilt, or torn down. Most users of our statistics have combined data on the availability of plumbing and condition as in accordance with a classification "substandard-not substandard" that was developed by the Federal housing agencies. By this classification, housing was substandard if it was "dilapidated," or, lacked or shared one or more of three plumbing facilities. The term "substandard" has not been used as a census classification of housing.

Since the 1960 census, much effort has gone into evaluating the census concepts and methods of rating the structural condition of housing, i.e., whether it was sound, deteriorating, or dilapidated. A summary and interpretation of the research carried out at the Bureau through mid-1967, was presented in Working Paper No. 25. This paper found that the 1960 statistics on condition were unreliable and inaccurate.

Concerning the rating of condition by enumerators this paper concluded that:

> "There does not appear to be any feasible method of improving the quality of enumerator ratings in a decennial census. This is a consequence of the ambiguities, nonoperational elements, and complexities of the rating process itself, as well as the biasing factors in the environment in which ratings have to be made."

Tests to determine the feasibility of direct self-enumeration of structural faults produced unsatisfactory results. It was recommended in that report, therefore, that:

> "The Bureau of the Census should direct its efforts to constructing measures of housing and neighborhood quality based on combinations of objectively defined characteristics."

It was also recommended that new measures be evaluated for "accuracy and validity . . . by comparison with measures based on the American Public Health Association (APHA) method . . . or some other maximum intensity method based on it."

II. Research Guidelines

A. Housing Advisory Committee guidelines.--Our research and discussions with the Census Bureau's Housing Advisory Committee and with the Task Force on Housing Quality, have established certain guidelines. These are:

1. A measure of housing quality must be national in application;
2. it must reflect with reasonable accuracy differences in housing quality between different areas.

B. Restrictions imposed by decennial census.--These guidelines must be related to the restrictions imposed by the use of the decennial census as the data collection vehicle.

First, the census depends upon the willingness and the ability of the population to answer the questions. Thus, the questionnaire must not impose an unreasonable burden on the respondent in terms of time; it should not deal with topics that appear to have no relevance or significance to the respondent; and it should not appear to be of an investigative or regulatory nature. Furthermore, the inquiries should request information that most respondents are able to provide.

Second, a decennial census of 70,000,000 housing units is of such a massive nature that tailor-made questionnaires for individual SMSA's or different geographic areas to reflect varying building codes, climate, population densities, or local needs, cannot be considered because of the overwhelming logistical, collection, and tabulation problems.

Third, the census is a household inquiry that obtains data on the housing unit and limited information on the structure. It is not a vehicle designed for directly collecting neighborhood information such as, land use, street utilization and design, air pollution, noise levels, the availability or quality of community facilities, or crime and health rates. Therefore, the characteristics used to identify poor quality housing pertain to the housing unit only and to the extent that these other factors are not considered we have only a partial measurement of the overall adequacy of housing. To put it another way, we are attempting to identify bad housing in terms of bad housing units and not in terms of neighborhoods or blighted areas.

C. Specifications for geographic comparability.-Keeping these limitations in mind we have been striving to find a method of identifying "bad" housing that will permit place to place and area to area comparisons that are not affected by—

1. Geographic differentials in methods of construction, climate, or location;

2. size of the community or the density of the population;

3. ethnic or socioeconomic characteristics as well as the attitudes and expectations of the occupants, respondents, or interviewers; and

4. elements of the immediate environment or neighborhood.

III. Proposed Housing Quality Measure for 1970

A. Components of measure.--The Census Bureau is considering a proposal to identify "bad" housing as consisting of two components:

1. Units lacking or sharing plumbing facilities, and

2. units with plumbing facilities but with high APHA penalty points.

The first group will be defined as it was in previous censuses. The identification of units in the second group will be based upon rent or value, and one or both of the following, heating equipment and kitchen facilities, in lieu of APHA inspections. (The evaluation of this identification process is discussed in part IV.)

B. Identification of high penalty point housing.--The identification of "bad" units with all plumbing facilities constitutes the major problem. Although we are not trying to duplicate counts of units based on the standards that were applied in 1960, we can place this problem in perspective if we examine the 1960 data. There were 10.6 million "substandard" units of which 91.5 percent, or 9.7 million units lacked plumbing facilities and 8.5 percent or 0.9 million units were dilapidated with all plumbing facilities. These 0.9 million accounted for about 1.5 percent of the total housing stock. In cities of over one million the proportion of the total "substandard" units which were "dilapidated with all plumbing facilities" was 21.7 percent as compared with 8.5 percent for the Nation as a whole. However, even in cities of that size such units accounted for less than 2 out of every 100.

The characteristics that we are proposing for use in 1970 to identify housing with plumbing as "bad" are:

1. Complete kitchen facilities.--Complete kitchen facilities are to include a stove or range, a mechanical refrigerator (not an icebox), and a kitchen sink, for the exclusive use of the occupants. These facilities need not be in one room.

2. Heating equipment.--Heating requirements and types of heating equipment in the United States vary between the North and the South and, therefore, we have tentatively established a dividing line based on a 3,000 heating degree day line.[1/] North of this line the following types of heating equipment are considered unsatisfactory: circulating, radiant, or room heaters, without flue or vent; fireplaces or stoves; portable room heaters of any kind; and none, (unit is not heated). South of this line, only fireplaces or stoves; portable room heaters of any kind; and none, (unit is not heated) are considered unsatisfactory. Heating equipment is not to be used as a criterion in the State of Hawaii.

3. Rent and value.--Basic to this proposal is the hypothesis that within a housing market, price generally indicates quality, i.e., the lower the price the poorer the housing. More specifically, in a market area where the price for some housing with plumbing is below the price for housing without plumbing, it is assumed that such units with plumbing have a sufficiently low position in the market place to indicate poor quality. However, in studying rent and value distributions it was found that in 1960, in many large cities, Negroes paid more for unsatisfactory housing than whites. Therefore, it is proposed that separate universes for Negro and white households be established in using the rent and value data.

Three methods were considered for applying rent and value as criteria for "bad" housing:

a. Rejected proposals.—

(1) We could use a flat dollar amount, for instance rents less than $60.00 or values less than $7,500. This idea was rejected because the quality of housing associated with a given amount varies greatly between geographic areas.

(2) We could use a flat percentile of all rented units across the country, e.g., the lowest 20th percentile in each city. This idea was rejected because it intimates that each SMSA has an equal proportion of "bad" rental housing, and thus, geographic comparisons would be invalid. The same reasoning pertains to the use of a flat percentile of value.

b. Proposed procedure.--The proposed procedure calls for tabulating rent and value distributions for units with and without complete plumbing facilities. Previously determined percentile cutoffs would be applied to the distributions of units without complete plumbing facilities. The dollar amount corresponding to these cutoffs would be applied to the distributions of units with complete plumbing facilities. This would be done for renter and owner units separately, by Negro and Non-Negro occupancy. Only four percentiles would be used across the entire United States. This procedure would be followed in approximately 400 "market areas." These would consist of each SMSA

1/One heating degree day is counted for every degree by which the mean daily outdoor temperature is less than 65°. The number of degree days for a locality represents the total for a year.

separately, plus the Non-SMSA portion of each State divided into three parts, namely:

(1) All cities of 25,000 to 50,000 population;

(2) all cities of 2,500 to 25,000 population; and

(3) all rural.

To illustrate, let us assume that the 67th percentile would be used nationwide for Negro-occupied renter units without all plumbing facilities. In SMSA "X" we would find that the dollar amount corresponding to the 67th percentile for these units is $54.00. We would then identify the Negro-occupied renter units with all plumbing facilities that rented for less than $54.00 and had either inadequate heat or lacked complete kitchen facilities as "bad" housing in SMSA "X." Similar steps would be taken to identify "bad" housing for Non-Negro-occupied renter units and for owner-occupied units.

IV. Current Research

To test this procedure and the hypotheses upon which it is based, we have underway at this time two studies.

A. American Public Health Association (APHA) rating of housing condition.--A sample of three hundred housing units with complete plumbing equipment was selected in each of three cities: Austin, Texas; Cleveland, Ohio; and San Francisco, California. The APHA rating of the structure and dwelling unit was obtained for about 800 of these units. Among the items included in this survey are access (from street, rear yard, or alley), rooms lacking an installed source of heat, an index of toilet condition, a report on infestation by rodents or other vermin, degree of deterioration in specified parts of the structure, and habitable rooms lacking closets. Deficiencies in the structure and the dwelling unit are expressed in penalty points; the greater the number of penalty points, the more severe the conditions revealed in the survey.

This study has three purposes. First, to determine whether units with high numbers of penalty points are also identified by rent or value in combination with type of heating equipment and/or availability of kitchen facilities. Second, to determine whether fixed percentiles of the rent and value distributions identify the same levels of penalty points in these three cities. Third, to gain some insight into the proportions of units in the United States that fall within various ranges of penalty points.

B. Housing Supplement to the Current Population Survey.--A housing supplement has been added to the Current Population Survey, for August 1968. Information is being collected from about 38,000 households on housing items that are used either directly as components of our proposed housing quality measure, or are required to assure accurate derivation of such a component. For about 19,000 of these units we will also obtain a rating of structural condition using 1960 concepts and methods.

The principal purpose of this survey is to obtain up-to-date data on the distribution of (1) heating equipment by categories proposed for the 1970 census; (2) rents; (3) values; and (4) presence of kitchen facilities as proposed in the 1970 census; for various classifications of geographic areas and specifically above and below the 3,000 heating degree day line.

The results of this survey will permit us to evaluate the impact of differences in classification of heating equipment above and below this line.

V. Conclusion

If the results of the APHA survey indicate that there is a satisfactory overlap between those units identified by the proposed objective criteria and those units identified by high penalty points, and if the August 1968 CPS Supplement shows there are sufficient housing units with the critical objective characteristics needed to identify "bad" housing units both above and below the 3,000 heating degree day line and in places of different sizes, we will conduct a large scale APHA survey in the spring of 1969 to define precisely the percentile cutoffs of rent and value to be used in the 1970 census.

If our current research indicates a low degree of mutual identification between the two systems, i.e., rent or value in combination with inadequate heating equipment, and APHA penalty points; or that the census items do not identify any reasonable number of "bad" units with plumbing in certain parts of the Nation, we will be forced to abandon this method for identifying housing units which have plumbing and also have high penalty points. We will then study the relationships between the individual census items, and levels of penalty points, to determine which objective characteristics will be shown in the city block reports and other publications. These data would not present a measure of housing quality as was done in the past but would provide the users with a broad spectrum of facts which, when combined with local knowledge, will assist in identifying areas of poor housing.

THE USES OF HOUSING QUALITY DATA IN PUBLIC POLICY FORMULATION

Frank S. Kristof, Housing and Development Administration, City of New York

The quality of the Nation's housing has been a public policy issue since the Housing Act of 1949 defining the nation's housing goals as requiring: "the elimination of substandard and other inadequate housing through the clearance of slums and blighted areas, and the realization as soon as feasible of the goal of a decent home and suitable living environment for every American family."[1/] Once the Congress mandated the elimination of "substandard" and "inadequate" housing as national policy, the definition of these terms became essential to a determination of the extent of Federal assistance required to achieve these goals. Similarly, since housing programs are undertaken by local communities, these measures were needed in order to ascertain the magnitude of a community's slum clearance and reconstruction programs. How many units require demolition, major rehabilitation or only moderate upgrading? Cnly when these questions are answerable can the level of resources required for a reasonable rate of progress be ascertained.

The second major need for this measure is to permit a community to determine the rate of progress it is or is not making in meeting its housing needs. Many bitter political struggles, local, state and national, have been waged on this issue. Today, nearly two decades after the Housing Act of 1949, the subject has lost none of its heat.

Part of the problem lies in the difficulty of finding quantitative measures for a qualitative concept. The technical problems associated with measurement of housing quality probably constitute the single most difficult task of the three decennial censuses of housing behind us and also in the forthcoming one. Since these facets of the subject are discussed by others, the point will not be belabored here. It is sufficient to say that this type of difficulty is generic to the social science field, particularly where the intervention of public policy substitutes for the operation of the market place. Consequently, the setting up of standards as a guide to public policy cannot be avoided even though one may quarrel with both their level and arbitrariness. That people's perception of what standards should be changes over time, only complicates an already difficult problem. But even this could be lived with if a fairly consistent measure of housing quality could be developed that would permit the scope of the housing job to be ascertained along with some definition of progress over time.

Another part of the disaffection over the measure of housing quality is that the texture of the housing problem has changed materially over the past two decades. When the Housing Act of 1949 was passed, the housing problem of America was set in a relatively simple framework of clearance of slums and their replacement with new housing as part of a program to make standard quality housing available to all families of the nation.

It was assumed that private investment would be used to rebuild the slums while low-income families would be rehoused by public housing. These assumptions proved to be not only over-simplifications lacking in validity but they failed to take into account existent social, demographic and economic forces whose implications in 1949 were not yet clearly perceived. Among these forces were the following:

1. The acceleration of Negro migration to the cities in the years since 1949, although this trend had been evident for at least a half century. From a scattered rural status in the nation's South, by 1966 Negroes had reached a point where 56 percent of their total population in the United States was concentrated in relatively small sections of most central cities and large areas of a few.

2. The simultaneous mass exodus of the white middle class to the suburbs. Again, migration to the suburbs is an historic American custom, neither new nor unique. The trend has been greatly accelerated in the past two decades by a combination of circumstances such as triumph of the automobile as the dominant means of transportation, the proliferation of roads and highways that opened up large new areas of residential development, and government mortgage insurance programs that encouraged a vast flow of private investment funds into suburban single-family home construction.

3. The decentralization of industry to outer areas. Although this occurred for technological and operating space reasons, it was encouraged and reinforced by the movement of population to the suburbs.

4. The negative impact of the above developments upon the economic base of the older cities that watched a steady outward drift of their manufacturing employment is a matter of record. Other consequences were a weakening of central business districts of such cities and an erosion of their real estate tax base as land and property values faded.

5. The wholesale exchange of better educated, more skilled and higher-income white

population for lower income, poorer educated and unskilled nonwhite population left the cities not only literally poorer but threw new educational, health service, and welfare burdens upon the cities. Thus the past twenty years have witnessed a major shift of Negro poverty from southern rural areas of the nation to the central city slums. In short, this represents a transfer of a major portion of the national poverty problem to communities that economically are poorly equipped to bear this burden.

The collective impact of the above developments was to make untenable the assumption that private investment funds would flow into the renewal of city slums. Such funds were far more readily attracted to the growing new suburban areas outside the cities. Furthermore, these investment areas entailed considerably less risk and trouble compared with the complicated and excruciating task of dealing, in the renewal process, with endless layers of municipal, state and federal governments that functioned sometimes in tandem, more often in conflict, and almost always sequentially. Time dimensions of the renewal process have become so great that only the youngest and most adventurous entrepreneurs now are willing to tackle this task. Only are such persons likely to live long enough to see the projects they have the courage to initiate evolve into completed, if not always profitable, ventures.

The second assumption--that public housing would take care of the housing of low-income families--also broke down. As the magnitude of the task expanded with the influx of low-income nonwhites into the cities, the popularity and even the acceptability of public housing declined as it became clear that this type of housing was developing into nonwhite, low-income ghettoes. Thus, today, aside from being only one facet of the problem of poverty, the housing difficulties of urban America have been compounded by the racial aspects of poverty.

The complexities involved in the regeneration of city slums are recognized clearly in the Model Cities provisions of the Housing Act of 1966 as a problem of restoring the total neighborhood. The concept includes replacement or provision of major facilities such as housing, public transportation, schools, hospitals, health centers and cultural facilities as well as financing educational, sanitation, social and protective services that have deteriorated seriously with the declining ability of municipal income to cope with the increased need and demand for these services. Finally, the need for development of employment opportunities was recognized as a major concomitant of such a program.

Along with acceptance of the concept of total neighborhood regeneration, went emphasis upon the preservation and rehabilitation of existing structures rather than their complete replacement. Thus the problem of measurement of the quality of housing assumes new and larger dimensions. Pressures for more accurate measures that will help a community assess the degree, as well as the magnitude, of its housing problem are multiplied as the complexities are recognized and more sophisticated approaches and resources become available.

The pressures come from two sides:

1. Statutory mandates of the Federal Government that local communities must meet specific planning standards to qualify for various types of renewal and housing funds.

2. Growing pressures of public policy makers within the local community that they be furnished with a more adequate data base for decision-making purposes.

The planning orientation of Federal housing legislation received further impetus in the Workable Program requirements of the Housing Act of 1954. During the 1960's, Community Renewal Program Funds provided support for development of objective criteria for the allocation of housing and renewal resources.

One of the outcomes of Federal legislation undoubtedly has been the influence on the professionalization of municipal housing and renewal efforts. On a larger front, one of the less well recognized facets of twelve years of administration under Mayor Robert F. Wagner is that major advances were made in the professionalization of the upper appointive echelons of municipal government in New York City.[2/] This trend, nurtured in the Wagner Administration, has been developed and extended over the past 2-1/2 years under Mayor John Lindsay. The massive reorganization of the Executive Departments of the City government has been accompanied by an influx of new, young blood that, if sustained and retained for any reasonable period of time, will provide a major infusion of new and talented personnel resources not experienced by the City since the depression of the 1930's drove into the security of civil service tens of thousands of young college-trained persons of that day. Supplementing the inflow of new talent has been the increased utilization, through new staff acquisitions and on a consultant basis, of modern analytical techniques such as systems analysis, operations research, program-planning-budgeting approaches pioneered by the Defense Department, the Budget Bureau and other sectors of the Federal Government.

The reorganization of City agencies into Administrations that combine similar activities resulted, in the housing and renewal field, in the consolidation under the Housing and Development Administration of seven major functional activities formerly carried out in four independent departments--urban renewal and moderate-cost housing development, relocation, management of City-owned properties in development areas, code enforcement, rent control and servicing of private new construction activity.3/ The Department of City Planning and the City's (low-rent) Housing Authority remain independent agencies although Edward Logue, in his report to the Mayor, had urged that these functions also be consolidated under the Housing and Development Administration. Departmental lines have been retained for operational purposes but for policy purposes, department heads now are deputies of the Administrator.

The potential for development of a housing program whose components would be consistent with one another has been greatly enhanced under the reorganization. Never before had the formal structure of government encouraged or even permitted examination of the inter-relationships of various housing and renewal activities other than through special investigations or ad hoc studies authorized by the Mayor. Prior to reorganization, for the analytical staff or even the Commissioner of one housing agency to comment upon the possible adverse effects of the activities of another housing related agency, would have been considered administrative interference and would have subjected the transgressor to rebuke or worse. Under such conditions, it is possible to obtain (after seven years) a Community Renewal Program report wherein the only passing reference made to rent control in New York City is that "rising costs (of operating old buildings) deter many landlords, especially those who own rent-controlled buildings, from making necessary improvements." On the subject of code enforcement, systematic area-wide cellar-to-roof inspections are regarded as a potential assist to preventive maintenance in relatively good areas while the Report notes that the most conspicuous result of full-scale punitive measures directed at landlords has been that "landlords have increasingly abandoned buildings."

Both the housing reorganization and the onrush of events have resulted in placing under critical scrutiny the relationship of rent control and code enforcement to the rate of deterioration and abandonment of structures in the City's housing inventory. In the brief period of the past six years, the City has witnessed a startling transformation in its housing situation. Between December 1962 and February 1965 available rental vacancies in the City nearly doubled--from 37,500 to 68,400 units. More significant is another new phenomenon--the sharp increase in vacant, boarded-up, or abandoned rent-controlled apartment structures, with or without occupants. As of December 1967, a computor run indicated that this figure had reached 2,000 buildings with 18,500 units plus an additional 700 buildings classified as rooming houses or other transient quarters that contained probably another 10,000 single-room units.4/

A field review of this list revealed that it was not complete. In areas where the investigation was carried out, an average of two vacant structures were found for each one listed. There is good reason to believe that the 2,700 buildings on the list more readily may total somewhere between 4,000 and 5,000 containing 40,000 to 50,000 units.5/

A development of this incredible magnitude in a City whose vacancy rate probably was the lowest in the Nation as recently as 1962 has created both a serious set of problems and opened up major new renewal possibilities. It should be made clear that the abandoned vacant buildings phenomenon is not a unique one--it is a frequent occurrence in cities with a declining population and an increasing housing supply. Philadelphia experienced this seven years ago; it also has occurred in Cleveland, Detroit, Baltimore and Newark. New York City, with the largest collection of on-going public and publicly-assisted housing programs in the Nation, has added about 200,000 such units since 1950: about half are low-rent and half are middle-income rental or cooperative developments. The private sector has contributed nearly 500,000 houses or apartments. The net addition to the City's housing supply (after deducting all losses) was about 500,000 units. Over the course of this period, meanwhile, the City's total population is estimated to have decreased by approximately 200,000 (from 7.8 million in 1950 to 7.6 million currently). Sooner or later, the cumulative effect of this continued injection of new housing into the City's supply during a period when its population was declining had to show up somewhere. During the early part of the present decade, the major impact was on the City's rooming house inventory. Our current estimates are that in excess of 45,000 rooming house units have disappeared from the market since 1960--primarily through demolition and to a lesser extent through conversion to apartments. Tens of thousands of Negro and Puerto Rican households have shifted from rooming houses into low-rent (controlled) apartments that have become available in many of the older sections of the City.

The climax to this process came when the adoption of a new and more restrictive Zoning Ordinance in December 1961 led to an outburst of new construction that jumped the rate of new units completed from an average of 31,000 in the 15 years 1947-1961 to 52,000 annually for the four years 1962-1965.

In the effort to "beat the Zoning Ordinance" builders placed into construction every buildable plot in the city that they owned or could acquire. The 4-year output of 209,000 new units resulted in a net addition of 186,000 units to the City's total inventory. Estimated net new household formation during this period equalled about 116,000 households which led to a gross addition to the city's vacancy supply of about 70,000 housing units. As a consequence of the chain of housing turnover that ensued, the new units gradually were absorbed and tens of thousands of vacancies began to appear in other sectors of the housing inventory, many in the form of vacant and abandoned buildings scattered throughout the city. In essence, a large part of the abandoned building phenomenon in New York City is associated directly with housing output substantially in excess of new household formation over the past two decades. For years, the City has been waiting for the day when the supply of housing would reach a stage wherein a sufficient surplus of housing would appear that would permit old and obsolete units to be drained off and replaced in an orderly renewal process. Unfortunately, the market does not work in such an orderly manner. First, these vacant buildings are scattered throughout the City. And in those sections of the city where they have been concentrated, the result included some severely destructive side-effects. In some areas where vacant structures became numerous, they constituted a hazard and danger to the residents of the neighborhood. In sections of the city, such as the South Bronx and in Brownsville-East New York in Brooklyn, where this phenomenon was accompanied by rapid change in the racial and ethnic composition of the neighborhood, the emptying structures along with the accompanying vandalism and criminal activity made the neighborhood literally unsafe to the older residents, with the result that entire blocks have been evacuated and many structurally sound buildings have been lost.

In one sense this may be regarded as a deplorable loss of useful low-rent housing. But where it occurred in areas that are being fled by their occupants, it also is a reflection of neighborhoods that no longer are functioning. From this standpoint, the Brownsville-East New York phenomenon may be regarded as an opportunity to clear and rebuild neighborhoods since the opportunity to rehabilitate and preserve them has been lost.

The Administration has taken this view and has used Federal funds to demolish, in Brooklyn alone, some 500 structures comprising about 2,500 units in the latter half of 1967. A large proportion of these were in Brownsville-East New York. It was ascertained from data supplied by the Bureau of the Census that 2/5 of the housing demolished with Federal funds in 1967 was classified as substandard in 1960. These demolished buildings represent the end of a market process that clearly demonstrates the functioning of "filtration" in periods of high output that exceeds normal market demand.

If this were the end of the story, it might be said that the city's problem is not overly serious. We have discovered to our dismay, however, that the abandoned building phenomenon is not limited to decaying areas and obsolete buildings. This process also has encompassed perfectly sound, solid, elevator-buildings built to modern standards in the 1920's. To some extent, such buildings may have been caught up in the weakening market in low-income areas, particularly those involved in rapid changes in racial and ethnic composition. Evidence exists, however, that some portion of the abandonment process may constitute much more than the normal market process described above. As a consequence, a full-scale investigation into this snow-balling development has led to investigation of concentrated code enforcement activities of the city as well as punitive rent-control actions. The inquiry currently under way is designed to ascertain to what extent a program of intensive code enforcement (cellar-to-roof inspection of every multi-family structure in a neighborhood) designed to induce owners to eliminate code violations, actually has had the reverse effect. Legislation permits the City's Rent Agency to reduce rents in buildings that have certain code violations. Serious questions have been raised as to whether this combination of activities has accelerated the abandonment process. Thus the correlation of cyclical code enforcement, rent reductions, and the subsequent abandonment of buildings is under intensive investigation. Since most owners of marginal buildings with controlled rents are faced with a negative cash flow situation in removing code violations, it appears that many do not find it worthwhile to continue operations when punitive action is undertaken.

In addition to the above investigation, the Rent Agency earlier had contracted with Professor George Sternlieb to conduct a major study of the cost of operating rent-controlled structures in an effort to ascertain definitively the overall operational economics of the rent-controlled sector.[6/] The Rand Corporation is conducting for us an even wider study of housing costs that covers all housing ranging from the poorest slum structures to luxury buildings. It is clear from the above that analytical approaches that cut across parochial departmental boundaries offer dramatic possibilities of coming to grips with problems whose political potency has been so great that they have remained untouched for many years beyond the point where rational investigation and correction was warranted.

The potential size of the deterioration problem in New York City that could lead to abandonment can be sketched in outline. Older rent-controlled buildings, owned free and clear, can be operated on a break-even cash flow basis of about $15.00 per room, (including a requirement for about $2.50 per room in taxes and about $1.00 for management) with no allowance for (1) return on investment, or (2) reserves for major improvements. An allowance for necessary reserves would increase break-even rents to $18.00 and if we wished to allow the owner a 15% return on the assessed value of the building, average rent per room would reach $20, a figure that could liquidate a reasonable mortgage structure.

In 1965, rents per room for rent-controlled structures in New York City were distributed as follows:

Rent per room	Number of units
All units	1,307,000 [a/]
Less than $10	27,700
$10 to $14	249,500
$15 to $19	242,100
$20 to $29	464,700
$30 to $49	240,500
$50 to $74	61,900
$75 or more	20,600

a/ omits 28,700 no cash rent units

Thus clearly in the danger zone are some 520,000 units with an aggregate of about 2.3 million rooms or nearly half of the 4.9 million rooms in the existing rent-controlled inventory. Crude calculations of the aggregate additional rental income required for proper maintenance of these buildings (using the $20 per room figure), or the implicit subsidy of renters being borne by a combination of landlords, the City and in the form of housing deterioration, amounts to approximately $155,000,000 or $300 per unit annually. This is a surprisingly small figure when we consider that it relates to the potential loss of two-fifths of our rent-controlled housing stock. This "deficit" also is a small figure next to an estimate of $300,000,000 as the City's annual welfare payments for housing. Incorporated in this estimate, however, are rent increases approximating 122 percent for the 2 percent of the rent-controlled units with rents below $10 per room; a 60 percent increase for 19 percent of the inventory with rents between $10 and $14 per room and a 14 percent increase for 18 percent of the inventory at rents of $15 to $19 per room. Hypothetically, if the City were to pick up, in the form of a City-financed rent subsidy, that portion of rent increases exceeding 20 percent of their income for all families whose rents were increased to $20 per room per month, it is doubtful that the City budget would be taxed by as much as $100,000,000 annually--which is pin money relative to the City's 1.36 billion dollar welfare budget.[7/]

Unfortunately, however, the above arithmetic omits a vital consideration. How many of the 520,000 apartments in the above calculation are in such condition that they require major rehabilitation or at least substantial capital replacements in order to put the buildings in sufficiently decent working order that a rent of $20 per room per month would properly maintain them?[8/] This question brings us to a full turn of the wheel. The objective data on the quality of the City's housing are so scanty and so incomplete that it is a source of continued embarrassment to City officials when they are asked, at least once a week, questions such as the following:

1. How many units in the City are so bad that they should be demolished?

2. How many units require (and warrant) complete rehabilitation (at an average of $10,000 per unit)?

3. How many units require only substantial improvements or equipment replacements to return them to useful life (range of $1,000 to $4,000 per unit)?

What is worse, the estimates that are available in the City, attributable to various agencies or individuals, are embarrassing in their range. For example, note the following estimates for New York City:

U.S. Bureau of the Census (1965)

1. Substandard units	214,000
2. All units that are deteriorating; plus sound units with shared plumbing facilities; plus all dilapidated units.	584,000

Community Renewal Program (1965 Report), City of New York

1. "Somewhere between 450,000 and 550,000 of the City's households cannot find housing of an appropriate size and rental to fit their needs."

2. "There are more than 800,000 dwelling units located in deteriorating areas where some form of governmental assistance is essential to restore a sound environment."

Frank S. Kristof--Housing Policy Goals and Turnover of Housing, Journal of the AIP, August, 1965.

Housing need (replacement of substandard units and standard units with 1.01 persons per room or more):

Jan. 1, 1960--385,000
Jan. 1, 1970--210,000

Census data alone, depending upon how one chooses to interpret them, provided in 1965 a minimum estimate of 214,000 substandard units to a maximum of 584,000 "unsound" units as interpreted by the Abrams Task Force Report to Mayor Lindsay in January 1966.[9] The Community Renewal Program report of 1965 with the estimate of "800,000 housing units located in deteriorated areas" is less helpful in arriving in magnitudes although the statement is a recognition of the neighborhood conception of the renewal problem.

On the other hand, my 1965 estimate of a decline of housing needs from a figure of 385,000 in 1960 to 210,000 in 1970 appears optimistic in face of recent developments suggesting that continued deterioration in rent-controlled structures containing over one-half million units plus an accelerating rate of abandonments, appears to be a clear and present danger. It is my judgement that, unless remedial steps are soon taken to provide sufficient incomes to owners of rent-controlled buildings to restore an almost lost incentive to the maintenance of their properties, the City will have a new public utility on its hands within the next decade--about 50,000 rent-controlled structures containing in excess of one-half million units in various degrees of obsolescence, decay and deterioration. A worse disaster than the withdrawal of private professional real estate resources from the management of low-rent properties cannot be conceived for New York or any other city with a large rental supply. The public sector is totally unprepared to cope with the magnitude of responsibilities implicit in such a transfer.

Certainly, the problem of a decaying housing stock is no monopoly of New York City. It is endemic to most of the older cities of America. George Sternlieb's pioneer work on The Tenement Landlord in Newark showed clearly that many forces were operative in creating the miserable slum conditions of an older central city. As the composition of the population in these areas changed, attitudes toward properties changed, born of a combination of inexperience, fright, stupidity and cupidity, on the part of tenants as well as property-owners. City services declined perceptibly with the advent of lower-income nonwhite families insufficiently sophisticated to develop political pressures for the maintenance of public services. Ownership shifted into weak, inexperienced or absentee hands and disinvestment in properties became the norm. The strongest attempts at maintenance of properties came from owner-occupants or professional real estate operators. As properties deteriorated, vacancies rose, rents fell, and the quality of tenants became more marginal. It should be noted that rent control played no part in the Newark story. Furthermore, rent control largely is irrelevant in areas of New York City that have reached similar states of dispair. Controlled rent levels generally are at, or at above, the market price in such locations and stringent code enforcement only compounds the problem.

The special problem for New York City is that areas of the city that contain fundamentally good housing where rents are substantially below market levels also appear to be subject to deterioration and ultimately, to abandonment of properties. To the extent that rent control is responsible for disinvestment and the withdrawal of professional real estate management from presently sound areas, the City is unnecessarily inviting disaster that a modification of the rent laws can avert.

For the rock-bottom areas of the city, only a massive infusion of public funds implicit in the Model Cities sections of the Housing Act of 1966 can help the situation. It will take a full-scale mobilization of all the public resources available, as well as the identification and involvement of the local residents, with these endeavors, to break the cycle of unemployment and low-incomes, poor health, poor education and other lacking public services. The dimensions of this task have not even begun to be perceived.

The task of reconstructing and rehabilitating the housing stock of these and less seriously affected areas requires a base of detailed and accurate information on condition and quality of the existing inventory. Housing oriented persons are well aware of the dissatisfaction, within the Bureau of the Census, of previous attempts to assess condition of housing in the effort to obtain some measure of quality of housing.

It is a frustrating fact, however, that if Census data on this subject are discarded, public officials will be left with absolutely no basis for arriving at a satisfactory determination of the quality of the city's housing inventory, to say nothing about satisfactory estimates of the number of units that require major or minor rehabilitation versus those that should be demolished. With the greater proclivity of public administrators to rely upon their planning and research staffs to generate such information as a guide to the allocation of housing and renewal resources, it becomes increasingly clear that the responsibility of the Bureau of the Census for improved housing quality information in 1970 is greater than ever before. That the problem is a thorny one is indisputable. But that it should be evaded or sidestepped at a time when such information is more vital than ever is inconceivable. Time is critically short for the work necessary to accomplish this end. It would be a most unfortunate failure if the Bureau of the Census did not meet the challenge necessary to give local public officials a usable tool to ascertain the magnitude of their housing task and to provide a reliable benchmark for a measure of progress.

1/ To be sure, the Housing Act of 1949 was not the beginning of housing as a public policy issue since many decades of local, state and federal legislation relative to many aspects of the subject preceded and was necessary to development of the political momentum that culminated in the 1949 Act.

2/ Mayor Wagner more than once expressed the view that he regarded this as his major contribution to New York City during his tenure. It is literally true that the majority of his political appointees to top levels of the municipal government had genuine familiarity and ability to deal with their areas of responsibility. He, furthermore, had occasion subsequently to commend Mayor Lindsay's efforts in accelerating this trend.

3/ The latter refers to examination of and approval of plans for new construction, issuance of building permits, construction inspections, and issuance of Certificates of Occupancy for completed residential structures.

4/ The 1961 figure did not include a count of vacant rooming houses and other transient quarters.

5/ A vexing aspect of abandoned buildings is that there is no single channel by which they come to the attention of the City. Some are noted when the owner obtains a certificate of eviction from the Rent Agency; others are recorded when complaints to the Code Enforcement Agency lead to their discovery, while inspectors in the course of their rounds note the existence of such structures and report them for the record. Another distressing facet of this story is that many owners have abandoned their buildings along with their occupants. Such owners have simply ceased to collect rents, literally disappearing from the scene.

6/ Professor Sternlieb's study of Newark's problem tenements (The Tenement Landlord) was a major contribution to knowledge in the analysis of owner and tenant attitudes and the economics of marginal tenement buildings.

7/ This is not quite the complete picture. If we were to assume that owners of all buildings with rental incomes greater than $20 per room were permitted to restructure their rent rolls to obtain an average of 15 percent more income (for which they would be deliriously happy) another $148, 000, 000 per annum would be added to the aggregate rent bill of tenants. Based upon the same supplemental City rent subsidy premise described above, it is doubtful that the City's welfare costs would rise by more than an additional $50 million per annum.

8/ Also relevant for public policy is how much would the capital expenditure be and how would it be financed and amortized.

9/ An "unsound" unit as used above is any unit that is not classified by the Census as "sound, with all plumbing facilities" for the exclusive use of the household.

DISCUSSION

Joseph Waksberg, U.S. Bureau of the Census

The two papers presented here display, by their juxtaposition, a frequent dilemma in social science: a situation in which there is an urgent need for information but only a limited ability to satisfy this need. On the one hand, Dr. Kristoff underscored the policy maker's requirement of knowing the magnitude and nature of his problem. On the other hand, both Mr. Young and Dr. Kristoff indicated the difficulties with previous attempts to find precise and accurate answers. Not all of these difficulties stem from the inability of the Census to train enumerators to recognize and identify poor housing. Even the notion of what constitutes poor quality housing is an unclear one, and as Dr. Kristoff has pointed out, different agencies reach different conclusions even when questions of measurement accuracy are not involved.

As both Federal and local agencies get involved in other social problems, many policy makers are becoming aware that they face a similar dilemma. In each case they have to make do with some simple, measurable items in the Census, to act as approximations for complex ideas that are abstract, intangible and tend to change their character over time.

Thus, the Poverty program has had to use as a basic measure of the extent of poverty and of progress being made, a fairly simple formula taking into account only the total income of each family and its size. This ignores such obvious aspects of a family's needs as assets, health, educational needs for children, age and future prospects, etc. Similarly, educational attainment in the Census is measured in terms of number of years of school completed, although it is by now quite clear that this is a very crude measure of the level of education actually attained. The needs for information on health are being served by simple counts of the number of persons who consider themselves as disabled, although the notion of disability is itself a complex one. Similar examples could be given in other fields.

In each case, the administrators have been as unhappy at not getting precise answers to their problems from the Census as Dr. Kristoff is.

However, even with the vast resources of a decennial census (the 1970 Census is expected to cost over 200 million dollars) it cannot do more than provide highlights and key indicators of the social and economic situation in the country. For a few subjects, an attempt to probe more deeply into their complexities is being developed by planning for a relatively small-size sample survey to follow soon after the census. The sample surveys will provide detailed data at a national or large area basis, which will assist the users in interpreting the cruder Census information published at the tract, city and county level. Perhaps it would be useful to consider the possibilities of a similar program for housing quality.

In listening to Dr. Kristoff describe the needs for housing replacement, which in New York City alone runs into hundreds of thousands of units (regardless of which of the different estimates is used) and the cost of which apparently will run into billions of dollars, I was struck by two contradictory thoughts. The first was the fact that if decisions of this scope are really to be based on hard information, then the administrators of these programs and housing analysts advising them had better realize that substantial increases in budgets for statistics and research are necessary. It is just not realistic to expect a handful of items in the Census to provide detailed information to justify the expenditure of billions of dollars.

The second thought was a skepticism as to whether information of such a precise nature is really necessary. I would infer from Dr. Kristoff's remarks that New York's problems are so vast that they can only be partially resolved within the foreseeable future. I assume it is the unavailability of money that is the limiting factor and not the lack of precise figures on the target population. Would New York's program really to any different if the target were 200,000 or 400,000? I am not implying that it is not important to have some measures of the scope of our problems and to know the extent to which we are making progress in solving them. I am questionning whether exact and precise measures are really necessary. If so, a vast increase in funds for their measurement will undoubtedly be necessary.

DISCUSSION

E. Everett Ashley, Health & Welfare Council

The papers prepared by Messrs. Kristof, Young, and Selove, provide a good point of departure for an exploration of "where we go from here," on the whole subject of housing quality and the broader and more important question of evaluating the urban environment. There are, I am sure, many people at the Bureau of the Census who sincerely wish that the Census had never been brought into the business of providing housing quality indicators and would be most happy if the whole matter could be swept under the rug. This can never be. As Kristof has observed, the need for the identification of the magnitude of the housing job faces every policy maker in the field.

As he points out, the absence of good agreed-upon criteria are a continuous source of embarrassment and confusion to housing and urban renewal administrators and to the legislative bodies which must provide the public funds so necessary to carry out the job. This is by no means the fault of the Census Bureau. Rather, one of the major causes of our present confusion results from the failure of either the Congress which put the statutes on the books or the Department of HUD and its predecessor which has the prime responsibility for funding slum clearance and low cost housing activities in the United States, to take the lead in setting minimum standards of housing quality. For more than 30 years the Congress of the United States has been adding to the body of legislation authorizing the expenditure of Federal funds for the elimination of slums and the promise of decent, safe, and sanitary housing for families of all incomes. In my mountain retreat where I prepared for this program I did not have access to the legislative history of housing and urban renewal in the United States. I cannot, therefore, recite the specific number of public laws which contain some reference to the question of "slums," "urban blight," "unstandard or inadequate housing." Through the years our legislative draftsmen, many of whom have been associated with HUD or its predecessor, and committees of the Congress, have been very free in their use of language referring to bad housing and bad environment. They have, unfortunately, been extremely niggardly in their specifications as to what they meant by these terms. In the Housing Act of 1949, the Congress went so far as to direct the Housing Administration to report to the Congress and the President on the state of the housing situation and the rate of progress being made in meeting the Nation's housing needs. Implicit in the meeting of housing needs is, of course, the elimination of bad housing. The stated national goal was the promise of a decent home in a suitable living environment for every American family. Yet the Congress gave not so much as a hint as to what a "decent" or for that matter an "indecent" house was nor what constituted a "suitable living environment."

Since the responsibility for reporting to Congress and the President on unmet housing needs is currently lodged with the Secretary of Housing and Urban Development and before the creation of the Department of Housing and Urban Development, with the Housing Administrator, one can well ask why has he not taken the lead in establishing national norms. Over the years it is true that the Secretary or his spokesmen have talked of the magnitude of national housing needs but nowhere, I repeat, nowhere will you find any official statement of what is or is not a "substandard" house or what is or is not a "slum."

The argument has been made that situations vary from place to place, that what is acceptable to some communities is unacceptable to others. Therefore, it has been argued, it should be the responsibility of localities to set their own norms. This contention used repeatedly over the years is in my judgement completely specious. Obviously a locality has the right and the privilege of setting its own standards of acceptability. Some plush bedroom community may elect to refuse to permit any house to remain in occupancy which does not have a garbage disposal, four electrical outlets in every room, a minimum of 2-1/2 baths, and a garage or carport adequate to shelter all automobiles owned by the occupants. Other communities may settle for housing which has at least a single bathroom, a weather repellent structure and no blatant violation of the National Underwriters Electric Code.

This does not in any way obviate the need for the Federal government through its Department of Housing and Urban Development being very specific about exactly what its criteria will be as far as the disbursement of Federal funds for urban renewal, for low cost housing, and for Model Cities. In the absence of such an explicit statement by HUD it is small wonder that this kind of confusion exists at the local level as Dr. Kristof points out in his paper. Although they are all under one overall management one might often come away with the impression that those who oversee the CRP program for HUD never speak to those in charge of public housing or of plain old urban renewal.

Both Kristof and Young make passing reference to but do not treat in depth an important facet of this whole housing quality problem, namely, the identification of slum areas and unsatisfactory neighborhoods as opposed to the identification merely of individual bad houses. As increasing emphasis is placed upon community renewal and upon the upgrading of vast neighborhoods through the Model Cities approach the great and urgent need is for some specific ground rules for measuring the quality of whole neighborhoods. How can the Secretary of Housing or the Congress, to say nothing of the localities themselves, tell whether a Model Cities effort is worth the candle when there are no agreed-upon criteria for identifying before, during, and after the fact, the quality of model neighborhoods?

The problem of environmental quality inherently is a tough one. I know because I have been involved in much of the experimentation which has been done on this subject over the past twenty-five years. I had been hopeful that perhaps from the present New Haven Use Study some clues would emerge of procedures which could be used for delineating slum or blighted areas by a merging of Census housing and population statistics with locally generated data to develop a mosaic which would differentiate neighborhoods of varying characteristics and quality. As things stand, HUD is failing to provide direction in this much needed research.

All of what I have said thus far, it seems to me, adds up to the need for the Secretary of Housing and Urban Development to take the initiative in setting what he regards as minimum standards which will make entire slum areas and their inadequate housing eligible for Federal assistance. Communities may at their option elect to enforce higher standards, but for purposes of Federal grants only those units and those neighborhoods below the HUD minimum would qualify for financial assistance.

This brings me to the really gritty part of the problem, the matter of objectively measuring neighborhood and structural quality in accordance with whatever standards HUD sets forth. Clearly it is not now nor has it ever been the responsibility of the Census to label houses as acceptable or unacceptable. Rather, as Young and Selove point out, it is its responsibility to provide measures which can be applied locally and nationally to identify bad housing according to the criteria set up by the user--most often either HUD or local agencies receiving HUD money.

The reluctance of HUD to speak its mind has not helped the situation at the Bureau of the Census. So long as HUD has not taken the lead and said what it needs for evaluating the urban environment and its housing, Messrs. Young and Selove and their colleagues in the Housing Division have been fighting a hard battle within the Bureau to defend the efforts to develop suitable measuring sticks for 1970.

Their task has been complicated by the decision of the Census to shift, in 1970, from an enumerator on the doorstep census to a mail census, i.e., one in which the respondent receives his questionnaire by mail, completes it himself and mails it back. Without visits to individual dwellings by trained enumerators, Census is obviously restricted in what can be collected that will be meaningful. I do not, however, share the doleful consensus of Working Paper No. 25 that, "there does not appear to be any feasible method of improving the quality of enumerator ratings in a decennial census." I have been involved in many tests and much experimentation along these lines. I realize the task is a tough one, but I submit that so far as developing useful national and SMSA benchmark data--not block by block detail--well trained enumerators could produce meaningful data even in a census type enumeration as a part of the components of change program. While improvements could be made in the techniques used by decennial census enumerators this will never fully meet the needs of housing data users. Instead, what is needed is the establishment of a regular recurring housing inventory survey which is oriented to the housing supply rather than to the household count. Certainly housing is important enough in the national economy to warrant such a regular survey just as a recurring survey of the labor force is an accepted piece of basic statistical fact finding. I would urge that HUD join forces with Census to secure the necessary funds to establish such a survey on a permanent basis.

As for the research which Young and Selove discuss I am inclined to believe they place greater confidence in APHA than is warranted. I fear it is merely a question of the grass in someone else's pasture looking greener. My personal observations of APHA rated houses in Louisville and some earlier experiences with an APHA rating of structures in Washington, leaves me a skeptic about what may come from that venture. What it seems to me is urgently needed is some intensive research to develop procedures which can minimize the errors which creep into present day APHA ratings. With the stake it has in this whole question of better measures of the urban environment I submit that this is an area into which HUD could well invest a substantial sum of its research money to enable Census and APHA to arrive at a more fool-proof procedure.

Perhaps it is because I have become mellow from sitting on my mountain top these past two months but I am inclined to support Young's proposal for testing the use of rent and value figures as criteria for "bad" housing. I suspect that it is going to be necessary, however, to go beyond a Negro—Non-Negro dichotomy. Puerto Rican, Mexican, and Indian families face the same kinds of market pressures as do the Negroes.

Turning for a moment to the work which is currently going on in New Haven, I feel that HUD will be most remiss if it fails to provide whatever financial support is necessary to assure utilization of the New Haven Use Study results for testing the feasibility of identifying not only "bad" houses but "bad" neighborhoods. The urgency of this is brought home in Dr. Kristof's paper when he points out the loss to the supply of good or at least salvageable units in bad neighborhoods.

In conclusion, let me say my hat is off to Art Young and Joe Selove for the work they have done and are proposing. If I do not share their enthusiasm or optimism for success along certain lines it is no lack of confidence in their skill. It is merely that after floundering around in this morass myself for a quarter of a century I have developed enough bruises and enough scar tissue to make me cautious. As time begins to run out on us it becomes increasingly urgent that by one device or another HUD, in its several manifestations, must be forced to the table to spell out with greater specifications than it has ever done up until now, precisely what it needs and how it will use the data it seeks. The need for such an explicit statement from HUD becomes all the more urgent as Census attempts to fend off efforts by misinformed Congressmen to curtail the size and scope of the 1970 census. The voice of HUD spoken loud and clear on the issue, with explicit examples of how the lack of data could seriously hamper the effective implementation of HUD programs, could give the Census the added muscle it needs in the present struggle. Somehow something seems to be a little bit out of focus when Congressman Olsen has to offer an extended defense of the 1970 census including the need for measuring housing quality and yet none of the papers I read indicated any statement of support from the Secretary of Housing and Urban Development.

DISCUSSION

Catherine E. Martini, National Association of Real Estate Boards

My comments are in three parts: First, a response to Arthur Young; second, a brief reaction to Frank Kristof, and, third, a general observation on the complexity of the subject.

Any technique of measurement, whether a verbal scale or a mechanical device, can be critically examined only in the light of the purpose the findings are to serve. Purpose will dictate the type of measuring instrument and the degree of precision appropriate.

Mr. Young has clearly, and in excellent detail, described the efforts of the Bureau of the Census, beginning in 1940, to identify housing that is a serious hazard to health and safety. The conclusions of Working Paper #25, that the statistics on condition are unreliable and inaccurate, combined with the data collection method planned for 1970, has led to exploratory studies designed to provide a new means of identifying deficient shelter.

While the 1960 method of measurement may be considered a failure as far as pinpointing individual substandard units, it has been thoroughly successful as a tool for delineating neighborhood areas that call for intensive and more specific measurement. Obviously too much was expected of a statistical approach that lacked precise definitions and depended on observers with limited training and diverse background.

The subject of housing quality is vast and complex. As a continuum from best to poorest, quality involves environment, structure, and use.

Oldest and most significant are the concepts and measurements expressed in building codes. Together with zoning, they provide regulation of the physical aspects of the structure and its relationship to nearby buildings.

Since only about 2 per cent of the housing supply comes into being within any one year, the maintenance and use of the existing inventory is of great concern. A Model Housing Law developed in 1914 was published in revised edition in 1920. In the preface Lawrence Veiller commented on the changing environment, making special reference to the automobile and the difficulties in adapting a model ordinance to local conditions. Light, ventilation, sanitation, fire protection, maintenance, and cleanliness, were covered. Environment, structure and use, as essential aspects of measurement, are clearly recognized.

Three years earlier (1917) a simple description of bad housing had been published by the Housing Committee of the National Association of Real Estate Boards. It read:

> "WHAT IS BAD HOUSING?
>
> "Houses having insufficient yards, making poorly lighted rooms.
>
> "Houses that are damp, unsanitary, filthy, in bad repair, exposed to undue fire peril, vermin infested, disease infected; having unclean yards and alleys; insufficient water supply, without toilet accommodations adequate to comfort, cleanliness and privacy; having overcrowded unventilated and dark rooms; without privacy, such houses are dangerous to the moral and physical health of the population of your city."

This definition of inadequacies was followed by a recommendation that local housing committees of real estate boards familiarize themselves with the Model Housing Law published by the Russell Sage Foundation and work with the National Housing Association headed by Lawrence Veiller.

In the interval between the collapse of the economy in the late '20's and the beginning of recovery at the end of the 1930's the Association moved forward in its concern regarding the condition of housing through activity centered in a Committee on Rehabilitation, which led to formation of the Urban Land Institute by a group of NAREB past presidents. Interest and basic concern of the Urban Land Institute at the time, was with deterioration of older residential areas that surrounded the monumental centers of cities. Its studies and activities led to the first proposal calling for Federal cooperation with municipalities for the acquisition of land in slum areas to be redeveloped by private enterprise. This was S.1163, introduced in the 78th Congress in 1943 by the late Senator Wagner at the request of the U.L.I.

If we are to be realistic, it is evident that the measurement of housing quality and public policy related to it should take two forms. The first is expressed in building codes and local housing ordinances with the highest possible standards of safety, maintenance, and sanitation, and severe personal penalties for violation. The second is the collection and use of Census data to identify areas that require attention, and for policy formulation related to Federal grants-in-aid and loans.

Since the Housing Act of 1954, which introduced the requirement of a local workable program, communities have been expected to prepare a comprehensive plan and to adopt adequate codes and ordinances. To a degree this will draw the two closer together and the proposed use of the American Public Health Association penalty points as a method of rating housing in the 1970 Census, will increase the similarities.

There will, however, always be vital differences.

The measurement of housing quality associated with local inspection, whether based on building or housing codes, is a continuous process, with personal penalties for non-compliance. The Census, as a data collection vehicle, measures housing quality only at wide intervals, and for individual units may be obsolete long before it is published. We can expect however, that the Census will continue to provide data for successfully delineating areas for more specific attention, a function that it has served well in the past.

This concludes my comments on Arthur Young's paper.

I disagree with Dr. Kristof that the measurement of housing quality is a public policy issue. Measurement is a scientific concept and should be totally separated from the action to be taken on the basis of its findings. In addition, policy implications mean to me much more than Federal legislation. They include the formulation of policy by the business sector, by State and municipal governments, and by interested private citizens as well. Public policy, in the basic sense, means attitude. I believe there is unanimity in the concern for better quality housing in a suitable environment for everyone.

An evaluation and goal for the Nation, centered on New York City data is, in many ways, irrelevant. By its own definition, New York City is atypical. Rent control has been maintained on the premise that the City is different from other metropolitan communities and non-metropolitan areas. The administration of control has flagrantly disregarded costs of maintenance and the role of maintenance in housing quality. The dismal study of vacant and abandoned structures testifies to the role of rent control as a destructive force, and points up the futility of using New York data as illustrative of other communities.

Further, New York City is atypical in the land use patterns through which it is developed, in the relationship of owner to renter occupancy; and in the percentage of tax-supported housing that has been built and in many respects disappointed its advocates who saw in it a solution to a broad spectrum of human ills.

To the degree that the phenomenon of abandonment has been observed in cities that do not have rent control, detailed area studies would be fruitful. The role of environment and of use in relationship to individual quality of a housing unit, might be studied for better understanding.

For any community, however, the suggestion that a census of housing should provide data indicating the degree of rehabilitation needed in dollar amount, is to fail to recognize the limitations of a census type of data collection and the purpose the results can be expected to serve. Estimating construction cost is a highly specialized function and requires training unlike that characteristic of an average census enumerator. If city officials are embarrassed by the scantiness of their information, local studies should be undertaken, tailored to the data needs of the particular city and the questions for which answers are sought.

This concludes my comments on Frank Kristof's paper.

In closing, I'd like to mention two general areas of concern regarding the measurement of housing quality and its policy implications. These are in addition to recognition that the subject of housing quality involves environment and use, as well as structure.

The first area of general concern revolves around the changing definition of a housing unit, depending on whether the shelter is occupied or vacant. Since a household and an occupied housing unit are synonymous by definition, ten per cent of the inventory is unoccupied.

The second is even more significant for the long run and, although touched on briefly by Dr. Kristof, should be given far broader attention. The need to improve the quality of housing of lower income families through subsidy is widely recognized. Virtually every legislative proposal arising out of present housing discussion that is directed at this problem, however, is expanded to include "middle" income groups which encompass the bulk of the population. Subsidies may be offered to some members of this group at times when fiscal and monetary policy lays a heavy burden on other members of the same group who are endeavoring unassisted to secure housing in the market place. We should admit that Federal housing subsidy on a vast scale is inevitably intertwined with monetary and fiscal policy and with the whole subject of horizontal tax justice.

IV

SURVEYING SOCIAL CHANGE IN COMMUNITIES

Chairman, MARGARET E. MARTIN, U. S. Office of Statistical Standards

SOME PROBLEMS IN INTERPRETING SURVEY OF RACIAL ATTITUDES IN 1968

Howard Schuman, University of Michigan

This paper will deal with two current issues in the study of racial attitudes. Both have come to be summarized in brief phrases, the first by the term "white racism," the second by the term "black separatism." The paper attempts to gain some perspective in interpreting the two phrases, looking more closely at what both phrases mean when attempts are made to operationalize them and gather survey interview data from the urban American public in 1968.

In both cases, time constitutes a crucial dimension. With regard to what is called "white racism," some important trend data are available, and together with current studies the trend results force a reexamination and suggest a conceptual reinterpretation of the phrase. In the case of what is called "black separatism," the emphasis must be not on the past but on projecting present results to the future. Only a little data are available on this, and I will in the present talk give the second issue somewhat less attention.

The phrase "white racism" used in the 1968 Report of the National Advisory Commission on Civil Disorders has had a considerable and controversial impact on public discussion. The Commission wrote:

> ...the most fundamental cause of the 1967 urban disorders is the racial attitude and behavior of white Americans toward black Americans....White racism is essentially responsible for the explosive mixture which has been accumulating in our cities since the end of World War II. (Report of the National Advisory Commission on Civil Disorders, p. 5)

To many black and some white Americans these words were a welcome focus on something that badly needs saying. To others the phrase "white racism" seemed overdrawn and unnecessarily loaded as a way of characterizing much of white opinion today. Social scientists are increasingly asked to evaluate the merits of the two sides of the argument, and indeed to declare whether most Americans are or are not "racists."

The findings to be reported here, however, indicate that some of the heated discussion over the issue is really irrelevant to the way the general public actually thinks about race. For the debate takes for granted the assumption of determinism common to almost all science, and revolves around which type of determinism the public holds. But much of the American public, it will be shown, does not accept or even consider this basic assumption at all. Blithely ignoring the logic of causal inquiry, much of the public operates on a premise of "free will," at least when thinking about Negro and white differences in status and achievement.

The phrase "white racism" is nowhere defined in the National Commission Report. But the term "racism" is generally taken to refer to the belief that there are clearly distinguishable human races; that these races differ not only in superficial physical characteristics, but also innately in important psychological traits; and finally that the differences are such that one race (almost always one's own) can be said to be superior to another.[1] More simply, "white racism" is the belief that white people are inherently superior to Negroes in significant ways, but that the reverse is not true.

Questions that can be said to tap "white racism" have been asked from time to time in national surveys over the past twenty-five years. The major finding of these surveys has been a dramatic decrease in beliefs in white racial superiority over Negroes. The most relevant and consistently measured topic has been white beliefs about racial differences in intelligence. In 1942 a National Opinion Research Center survey asked respondents: "In general, do you think Negroes are as intelligent as white people--that is, can they learn just as well if they are given the same education?" Only 42 percent of a national sample of white Americans said they believed Negroes to have the same intelligence as whites. Later surveys, however, showed a continuing rise in the belief in equal intelligence, so that by 1956, 78 percent of an NORC national sample answered the same question in the affirmative. The percentage seemed to stabilize at that point and more recent surveys have continued to show that about four out of five white Americans reject the notion that white people are born with higher mental capacity than Negroes.[2]

The comparatively rapid decrease in "racist" beliefs in this key area, the relatively small proportion of people who still hold such beliefs, and the fact that the hold-outs tend to come disproportionately from the old South, all suggest that racism--at least in the more open forms that can be measured in surveys--is a minor and disappearing phenomenon in this country. This, of course, implies little about the disappearance of discrimination or hostility toward Negroes or about other aspects of inequality in America. It merely indicates that attempts to buttress anti-Negro feelings with beliefs about biological racial inferiority no longer carry much weight with the white American public.

It is natural for social scientists viewing these trends to see them as indicative not only of the disappearance of "racist" beliefs, but also as an equally reliable sign of the acceptance by a growing proportion of the white population of one or more of the available contemporary environmental explanations of the Negro's disadvantaged status and achievement in America. It is easy to do this because this is just what has happened in social science itself. The type of deterministic assumptions that played such a large role in American social science in the early part of the century--beliefs in psychogenetic racial differences--has gradually been replaced over the last decades by explanations geared to environmental determinism. Some of these environmental explanations focus on what has been most obvious in the traditional American racial structure: segregation, discrimination, and the domination of Negroes by white power. More and more, social scientists have also looked to cultural and culturally induced psychological phenomena, such as the burden that lower class or rural background places on ability to compete for urban middle class rewards; the assumed disruptive effects of family instability and lack of successful male models; the disabling experience of growing up as a minority in a society where one's ethnic identity is both permanently fixed and negatively evaluated by a large part of the majority. Whatever the particular environmental theory, however, the important point is that an explanatory social science must look for causal variables that are independent of, yet can be said to produce, the "facts" that need explaining.

The problem is that the projection on to the general public of the logic of science leads to paradoxical results. For it is clear that although most of the American public reject "racist" beliefs, they do not emphasize environmental explanations of racial differences with the same fervor as do social scientists. The following question (Table 1) was asked in early 1968 of a probability sample of 2,584 white Americans in 15 major American cities.[3] The results show that more than half the sample believe that the inferior economic and educational status of Negroes is due mainly to Negroes themselves. Only 19 percent place the blame mainly on discrimination.[4] It is interesting to note that 4 percent of the sample denied the initial assertion of the question, claiming that in their city Negroes have jobs, education, and housing equal to or better than that of whites. This serves as another indication of how misleading it is for social scientists to assume knowledge and acceptance by the general public of social science findings--in this case descriptive findings rather than explanatory ones.

The term discrimination was used here as a simple way of representing clear environmental explanations, but it may have failed to provide sufficient opportunity for other environmental views, such as stress on Negro lower class background. However, a follow-up question (discussed further below) encouraged respondents to explain their ideas in their own words. Another 18 percent gave some sort of apparent environmental explanation, the major variant of which was mention of lower education on the part of Negroes. Of course, lower education by Negroes was already built into the question as part of the problem to be explained, and it may well be that many of the respondents giving this answer would attribute the educational deficiencies to Negroes themselves rather than to lack of opportunities. Still, making the maximum assumption of environmental emphasis here, and adding the 19 percent who mentioned discrimination explicitly, we find only some 37 percent of the sample attributing Negro disadvantage to causes outside Negroes themselves. More than half the sample place the responsibility for Negro disadvantage mainly or entirely on Negroes.

Table 1

"On the average, Negroes in this city have worse jobs, education, and housing than white people. Do you think this is due mainly to Negroes having been discriminated against, or mainly due to something about Negroes themselves?"

	Percents
Mainly due to discrimination	19
Mainly due to Negroes themselves	54
A mixture of both	19
Denied Negroes have worse jobs, education, and housing--refused question	4
Don't know	4
	100
	(2,584)

The results reported in Table 1 appear at first to contradict the trends reviewed earlier which showed a sharp drop in the tendency to attribute Negro lack of achievement to racial inferiority in intelligence. If Negro problems are attributed mainly to "something about Negroes themselves," does not this imply a "racist" explanation? The answer may be "yes" to the scientific determinist, but it is not necessarily yes to the general public.

The situation is considerably clarified by follow-up questions which we asked of the 73 percent of the sample that attributed lower Negro achievement to Negroes themselves or to a mixture of Negroes themselves and discrimination. We inquired first, "What is it about Negroes themselves that makes them have worse jobs, education, and housing?", and recorded the responses *verbatim*. No matter what the answer, we then asked: "Do you think Negroes are just born that way and can't be changed, or that changes in the Negro are possible?" Skipping over the free answer question for the moment, we found to our surprise that whatever the faults Negroes were seen as having, only 8 percent of the respondents saw these limitations as inborn and unchangeable, while 88 percent believed "changes in the Negro are possible." (The remaining 4 percent answered don't know.)

We thus arrive at a situation where we find that a considerable portion of the white urban population believes that the source of Negro hardships lies within Negroes themselves, but deny that these sources are inborn and unchangeable. The white public appears simultaneously to accept and to reject "racist" beliefs.

The resolution of the paradox is suggested by the free answers to the question: "What is it about Negroes themselves that makes them have worse jobs, education, and housing?" These answers were coded into the most meaningful categories inherent in the data, and the results are shown in Table 2. Only 8 percent of those asked the question speak in terms that imply or strongly suggest biological or genetic differences between Negroes and whites, and the number mentioning low intelligence as such is even smaller. This certainly does not contradict the NORC trend data presented earlier, but only accentuates it. Answers that lean in an environmental direction are given by a quarter of the sample, as noted earlier. By far the largest category of response, however, does not point in either a genetic or an environmental direction, but is best termed "lack of motivation"; 47 percent of those attributing Negro problems to Negroes themselves give such a response clearly, and another 10 percent offer related responses having much the same implication. Some examples from the interviews are as follows:

> "Well, they don't try to better themselves. I've come up through the ranks. I've worked at just about everything. And now I'm at a job where I'm happy and just about making top money. And they can do the same. Get out and look."
>
> "They have the same advantages the whites have but don't use them. They quit school. They quit work."
>
> "They pity themselves too much. We have Negro friends from the service, one is a hard worker and he has made something of himself. Many don't try to better themselves."

The proportion of responses like these to the proportion of responses focusing on intellectual or other types of lack of capacity is on the order of seven to one.

We now have three interlocking clues to what the majority of white Americans who are critical of Negroes see to be the main cause of disadvantaged black status in the United States. First, the cause is perceived to lie mainly within Negroes themselves, rather than coming from external constraints imposed by American social structure or by the prejudice of white Americans. Second, this interior cause is seen as a matter of motivation or will, not as a matter of capacity or ability. And third, it is not seen as either immutable or ineluctable, but rather as something that can readily be changed.

Changed by whom? For the white general public, the implicit answer is: by individual Negroes themselves! If that seems hard for the reader to fathom, it is only because the reader (like the writer) is a confirmed determinist, at least on race. But a great many Americans apparently are not. What they evidently believe in is a naive form of free will. Negroes can get ahead at any time, so the public thinks, simply by setting their sights higher and putting their shoulders to the wheel. The religious aspects of free will are rarely mentioned, to be sure, but free will is what the general public takes to be an explanation of how individual men and entire ethnic groups can, do, and should achieve success in America.

Table 2

Explanations of What It is "About Negroes" that Leads to Their Disadvantaged Status (Asked only of those replying "mainly Negroes themselves" or "mixture" to the question in Table 1)

	Percents
Responses that suggest genetic explanations of Negro disadvantage (e.g., "low mental ability," "low morals")	8
Responses that suggest environmental explanations (other than discrimination) of Negro disadvantage (e.g., "lack of education," "poverty cycle")	25
Responses that suggest lack of motivation as explanation of Negro disadvantage--no indication of genetic or environmental cause. (See text for examples)	57
Don't know, Not ascertained*	11
	100
	(1,886)

*The majority of these "not ascertained" responses were from persons who had answered "mixture" to the first question and who then answered the follow-up by discussing discrimination rather than what it is about Negroes themselves. This was not an answer to the question and is here treated as "not ascertained."

There is really nothing surprising in this public commitment to free will. It must in some form be built into every society, since elders and rulers usually feel it necessary to impress upon children and citizens the responsibility of the latter for their own actions. It may be that a person fails to live up to an important social norm only because of the way he was brought up or only because of the way his endocrine system functions; yet others in the society will not wish him to attribute his deviant behavior too easily to such causes. They will want him to hold himself responsible for his actions and to believe that he can change if he wishes to and tries hard enough.

Beyond this universal social need to hold individuals responsible for their actions, in America the emphasis on free will has an additional and very powerful cultural source in the belief that each immigrant group has started at the bottom and has proceeded by ambition and effort to work its way toward the top. Few things are more celebrated in our society. The second, third, or n-th generation descendent of immigrants is usually ready to recount vivid tales of ancestral initiative and industry. Moreover, told that Negroes have come from unskilled backgrounds, lack capital and connections, face prejudice and discrimination, many a white American will assure the teller this was all true of his own parents as well, or of his grandparents, or of at least an uncle or two. He will point out that despite tremendous obstacles his forebearers succeeded in America, and that that is exactly why they now have their house in the suburbs, their children in college, and the respect of their neighbors. He will assure the listener that Negroes can do as much at any point if only they exert the effort.

Such a reply will be deeply dissatisfying to the sophisticated social scientist. Even if he accepts motivation as somehow a major problem for Negroes, he will want to investigate why there is this motivational difference between black Americans and white Americans. He may also be so astonished at the apparent naivete of general opinion that he formulates more survey questions for the public. By pushing a good deal he may force some respondents to assert a genetic-like explanation, others to opt for a family structure explanation ("I guess it's the way their mothers and fathers brought them up"), and so forth. Yet these responses to probes will be given mainly to satisfy the pressure of the interviewer, not because the average respondent himself feels an explanation which assumes that human beings have "free will" leaves anything to be desired.

Putting together the results we have reviewed, it becomes clear that much of the public not only does not think in scientific terms about race; it does not even think in pseudoscientific terms. The general public does not look for deeper causes of Negro disadvantage because it sees these disadvantages as easily ended at any time by the very people suffering from them. Arguments over types of determinism are really irrelevant to this substantial part of public opinion, for it feels quite comfortable in thinking about race in the same simple free will terms that it uses in thinking about individuals: those who really want to get ahead can do so.

Turning to the issue of "black separatism," the results of our survey in 15 cities--in this case the sample consists of 2,814 interviews with Negroes--also challenge the applicability of the phrase. A clear-cut finding from our study is that the loud and often eloquent talk of some black separatist leaders is not very representative of the general Negro population in these 15 cities. Questions intended to tap total rejection of white society produce rejecting answers by only five to ten percent of the black sample. Nor is this simply a negative finding: questions on residential choice indicate that about 85 percent of the Negro population in these 15 cities have a positive preference for a mixed neighborhood or else claim that the race of their neighbors makes no difference to them one way or the other. When we focus on more restricted and currently very heated issues such as black control of schools and stores in black ghetto neighborhoods, we continue to find that the overwhelming majority of Negro respondents are opposed to the introduction of racial criteria into decision making. Most Negroes consistently apply principles of non-discrimination in these areas just as they do in matters where it is more obviously to their advantage. The proportion favoring black self-rule is, to be sure, a little higher on these questions, but it still constitutes at most 15 percent of the population.

Now all these percentages are very small. In an election poll or in a survey dealing with legislation before Congress, a candidate or an issue with so little support as black separatism has at present, would not expect to get very far. It is certainly useful to indicate this clearly. But at the same time, it is difficult to avoid the feeling that a focus on the 80 percent or 90 percent of the population that holds to the goals of integration is to miss something very important in the current and future dynamics of race in America. Therefore, we have tried to give more emphasis to what must be called the deviant cases in this area.

One simple way to do this is to translate sample percentages into population numbers. We estimate the Negro population in the fifteen cities to have been about 3,330,100 in early 1968. Thus when we speak of nearly ten percent of the black population rejecting white American society, we are speaking of nearly a quarter of a million individuals. This is a large number of disenchanted people, and the fact that it is concentrated in ghetto areas probably accentuates its influence by promoting communication and association among such individuals and by providing them with easy access to just the audience they wish to reach. Thus even 10 percent cannot be discounted if we are interested in predicting the future course of events, rather than in simply describing the present.

We feel fairly certain that this minority within a minority is growing, but at present we have no adequate change data to check this. It is possible, though hazardous, to use age as an indirect measure of change. If we do this we find a clear monotonic age trend in the data, with the youngest cohort showing the most change. Our 16 to 19 age category registers generally about 10 percent higher in separatist beliefs of almost all kinds than the sample average. Presumably the higher rate of separatism among the young indicates ongoing change and one can attempt to extrapolate the age curve to future cohorts. The assumptions become so shaky that we have not attempted to do this in any systematic fashion. Moreover, we must admit that the age differences are not as great as we had anticipated, and only extreme assumptions of recent acceleration lead to a projection from our data of rapid change toward separatism by even the entire younger generation of Negroes, let alone the Negro population as a whole.

There is one other finding of a purely negative sort that has led us to speculate about the future importance of the separatist subsample. Separatist responses do not show much association with common indicators of socio-economic status. While this makes their antecedents more difficult to disentangle, the fact that both high and low educated persons are well represented among separatist thinkers suggests two consequences. First, it means that this is not a movement only of the lower class, but that it can provide from its own ranks an educated elite for leadership and for the development of ideology. But second, if the adherents came _only_ from the upper SES levels they might well be too detached from the Negro masses to have much effect, becoming simply a new intellectual class or the kind of bourgeoisie that have run most large Negro organizations. The lack of relation to SES indicates that the movement--if one can call it that--has appeal and roots along the entire economic and educational ladder.

We are thus left with a feeling that black separatism has a future, even though its present is not very striking. This seems to be an area where the survey analyst must put together his time-bound data and his sense of what is happening to change the very source of his data. Rather than use the data only to emphasize the extent to which popular views of change are exaggerated, modesty impels one to qualify results to take account of emerging change that is indeed not yet apparent to the eye of the survey analyst.

REFERENCES

[1] A typical dictionary definition is found in _The Random House Dictionary of the English Language_, 1967, p. 1184: "a belief that human races have distinctive character that determine their respective cultures, usually involving the idea that one's own race is superior and has the right to rule others."

A recent definition offered by a sociologist writing on the topic is even more explicit: "Racism is any set of beliefs that organic, genetically transmitted differences (whether real or imagined) between human groups are intrinsically associated with the presence or the absence of certain socially relevant abilities or characteristics, hence that such differences are a legitimate basis of invidious distinctions between groups socially defined as races." _Race_ and _Racism_: A Comparative Perspective. New York: John Wiley & Sons, Inc., 1967.

[2] These figures are reported in Mildred A. Schwartz, _Trends in White Attitudes Toward Negroes_, Chicago: National Opinion Research Center, 1967.

[3] This and related questions and data reported below are from a study of "Racial Attitudes in Fifteen American Cities," directed by Angus Campbell and Howard Schuman. A preliminary report of the study appears in _Supplemental Studies for the National Advisory Commission on Civil Disorders_, Washington, D. C.: U.S. Government Printing Office, July 1968. The samples discussed here are of the white population, ages 16-69, in the combined 15 cities. The cities are: Baltimore, Boston, Brooklyn, Chicago, Cincinnati, Cleveland, Detroit, Gary, Milwaukee, Newark, Philadelphia, Pittsburgh, San Francisco, St. Louis, Washington. Results from two suburban areas (around Cleveland and Detroit) are essentially the same.

[4] Somewhat similar results are reported from a recent Gallup Opinion Survey with a national sample. The Gallup question reads: "Who do you think is _more_ to blame for the present conditions in which Negroes find themselves--white people, or Negroes themselves?" Only 23 percent of the white population blamed "itself"; 58 percent blamed "Negroes themselves"; and 23 percent had no opinion. _Gallup Opinion Index_, July 1968.

COMPARATIVE COMMUNITY STUDIES WITH LARGE N'S [1]

Robert L. Crain and Peter H. Rossi
Department of Social Relations, The Johns Hopkins University

I: Introduction:

The urban crisis is the major domestic issue of our times. It fills the front pages of our daily newspapers and provides materials for television documentaries. It is one of the central issues of the current presidential campaign. And, according to Gallup, the urban crisis in all its connotations ranks in the eyes of the American public with the Vietnam war as a foremost problem facing the United States today.

Partly as a consequence, urban studies have been restored to a position of prominence in sociology, political science and economics, a peak of popularity which it has not enjoyed for several decades. Undergraduate and graduate courses in urban studies are extremely popular. Research interest has also increased as social scientists have turned to the task of making themselves relevant and as funds have become available.

The urban crisis covers a variety of specific disorders ranging from traffic through education to race relations. Whatever catalogue of specific urban problems one would draw up to exhaust the meaning of the term "urban crisis", it is clear that at the core of the definition would be the related problems of poverty and race relations. In one sense, it is difficult to understand why poverty and race relations should be considered an urban problem when both problems are probably more serious in rural areas. In another sense, it is perfectly understandable why the term "urban crisis" has become virtually a euphemism for poverty and race relations because it is in the urban areas where the action is taking place. Poverty and race relations lie at the heart of the urban crisis because that is where the black poor are in ferment and as a consequence that is where the attention of our public officials and mass media is focussed.

A goodly portion of urban research is also concerned with the action -- how decisions are made, the course of political negotiations among public officials, civic leaders and black leaders, how resources are mobilized and distributed to meet the crisis and how the policies adopted are affecting the outcome of the urban crisis.

Research into community decision-making is just now emerging from the primitive stage of case studies of individual cities and specific decisions. To be sure, we have learned a great deal from the insights of sensitive researchers and from the concepts and propositions which have emerged from such case studies. However, at this point we have arrived at an impass confronted with contradictory findings from the many case studies and as yet unable to allocate the variance in findings to differences among researchers, methods or cities. This small crisis in urban research is widely recognized among researchers who are at the heart of the problem. However, few are willing to live up to the implications of this impass; the necessity for large scale comparative studies.

The main obstacle to facing up to this necessity lies in very heavy committment that a researcher would have to make to carry on comparative studies of communities using conventional research methodology. The few comparative studies that have been made so far (e.g. Agger, Goldrich and Swanson)[2] have taken years to accomplish and are expensive both in terms of resources and time, and few have either the funds or the patience to carry them out.

This paper is concerned with giving an example of how community studies involving relatively large numbers of communities can be undertaken, provided that one is willing to make certain sacrifices, trading off richness of data for large numbers of cases. Note that this strategy is precisely that which supports the use of sample surveys as opposed to intensive case studies of individuals. In other words, what we are proposing and illustrating in this paper is an analogue to the sample survey of households applied to the study of communities.

Properly worked out, this strategy should make it possible to conduct comparative community studies with large N's at unit costs in terms of time and resources considerably below that involved in the usual case study. It is a strategy which involves certain risks: It may mean settling for relatively crude measurements on any one particular community in favor of systematic measurements on a large number of communities. It may also involve restricting the number of variables measured to a relatively small number, increasing the risk that the variable(s) that may turn out to be critical have not been measured at all.

We feel that the advantages of large N comparative data justify the risks involved. First of all, such a strategy eliminates the tedious arguments over what are the "facts" concerning American cities. Floyd Hunter[3] writes that Atlanta, Georgia has a power structuve in which a handful of wealthy businessmen make all the important decisions and therefore cities are undemocratic. In contrast, Robert Dahl[4] writes that New Haven has a pluralistic decision-making apparatus and that therefore cities are healthily democratic. It is difficult to know whether these two authors disagree because their biases are different, because their methods differ or because the two cities they studied were different.

Secondly, because the universe of cities in the United States is not very large, the study of large samples of communities will lead inevitably to the accumulation of data over time about many cities making it possible to make statements concerning historical trends. We have proposed to maximize this feature by establishing a "permanent" sample of communities, representing a probability sample of cities in the country, in each of which a data collection apparatus would be set up making it possible to conduct researches whose results would accumulate over time representing eventually an archive of time trends.

Finally, the replication implicit in large N comparative community studies means that researchers will tend to converge on a common set of concepts and operational measures. We can therefore expect that the constrictions on speculation imposed by having relatively hard data will make it possible to move the conceptual development of the field of urban studies forward faster.

The viewpoint expressed in this paper is not unique to the authors. The researches of Oliver Williams, Heniz Eulau, Michael Aiken, Sidney Verba and a number of others all represent illustrations of this new approach. Our unique contribution has been to propose and set up an institutional device -- The Permanent Community Sample -- to make large N comparative community studies easier to conduct.

In this paper, we will draw upon data from a prototype large N community study to illustrate the nature and importance of the contributions these comparative studies can and will make. The sample concerns what has come to be a critical problem in the study of cities: Who has power?

II: The Problem of Community Power:

Our approach to the problem of community power is to regard power as residing in important sectors of the community rather than in individuals. Whether or not one starts with the attribution of power or influence to specific individuals, in the end the critical questions end up to be which of the major groups in a community wields more influence on the raising and their settlement of issues. Indeed, the controversy over community power mainly centers around whether elected political officials have more or less influence on the outcome of decisions than do the owners or managers of large business enterprises. The main difficulty with most studies of this problem is that they have relied upon informants to decide which group has more power, not recognizing that this is a problem in the analysis of facts rather than in simply knowing what the facts are. In other words, it is a kind of apples and oranges comparison.

Our concern here will be with three sectors of the community -- business and civic leaders; public officials and political party officials; and the general citizenry. Because the decision-making roles of each of the three groups are so different in nature, it is not possible to make sensible statements concerning which of the three has "more power" than the others. For example, in a very real sense, the citizenry have the final say since they have the power to replace public officials through the electoral process, but in a day-to-day sense the citizens are not well enough organized ordinarily to make much of an impact. Similarly, public officials by virtue of the fact that many decisions by law are left to them to make, can easily be shown to make most (but not necessarily the most important) of the public decisions in any community. In some cities this means that the dominant political party can set overall policy and enforce it through its control over elected officials and through them the municipal bureaucracy. Where elected public officials and party leaders lack power it is because they cannot defend themselves against the risk of being defeated in an election or because they lack the courage to run that risk.

In this rather ambiguous battlefield the civic leader and business man frequently emerges as the dark horse who winds up with most of the winnings. By playing the civic-leader game -- by contributing time, money, prestige and technical knowhow -- civic leaders drawn from the business sector of the community may often be able to take de facto control over decision-making.

These three groups are ordinarily engaged in a long and continual battle for favorable balances of power. The city becomes in effect a territory divided into three nations. But whenever a small battle is fought over the details of how much of one decision belongs to which group, it is not seen as part of a war. Everyone accepts the boundary lines of decision-making as natural and right when they represent the results of battles fought long ago and which have divided the territory into traditional "turfs".

For these reasons we can anticipate that the structural similarities among American communities will insure that each of the three major groups will be playing similar roles in each community but that the boundaries of traditional jurisdictions will vary from place to place. It is therefore difficult to phrase the question in terms of who has power over whom in each city, but it is easier to compare how much leeway each group has in each city. In other words, we can more easily make statements which rank communities in the extent to which business and civic leaders play important roles than we can make statements indicating the extent to which business and civic leaders predominate in decision-making over elected public officials.

III: Measuring Participation in Decision Making:

There are many problems in the comparative study of communities where it does not look as if it will be possible without very drastic innovations in data-collection techniques to conduct studies in a large number of communities easily and inexpensively. For example, if one is concerned about the way in which differences in public attitudes towards education relates to per capita expenditures for education, it is difficult to avoid taking sample surveys within each city, a process which rapidly adds up to a very expensive operation for even a small set of cities. If one were willing to settle for samples of size 100 (an unreasonably small size in the eyes of most researchers) in each of 100 cities, the interviewing task would be larger than all but the largest of survey operations could sustain and the costs would be correspondingly very high.

But, there are many problems in the study of which it is appropriate to consider using very small numbers of highly selected respondents in each city to provide information adequate to characterize that city. For example, if one is concerned with the policies followed by central institutions, e.g. the police department or public school system, it is appropriate to collect such data from a small handful of persons who are in a position to know about the issue in question. Thus a study of school board decisions concerning desegregation or the selection of a new school superintendent can be conducted sensibly by interviewing key persons involved -- the school superintendent and members of the school board. In short, when we are concerned with the organizational life of the community in circumstances wherein knowledge is concentrated in the hands of a few, we can more easily envisage how very large N comparative community studies can be undertaken without courting bankruptcy. The example given in this paper is one of the cases in point.

The data upon which this paper is based were generated by using a small set of persons as informants in each of fifty-one cities. In each city, we picked eight respondents, each selected because his position in the community provided him with particularly intimate knowledge concerning the decision-making processes of that community. Each informant was interviewed using a structured questionnaire by skilled interviewers of the National Opinion Research Center of the University of Chicago. Structured questionnaires were used to provide comparability across communities and were carefully pretested to insure that they neither offended the informant by being too simpleminded nor missed important information by being too crude to pick up local nuances. Each type of informant received a questionnaire tailored to that informant's particular sphere of competence although questions common to all respondents formed the core of the questionnaires used.

The eight informants selected were as follows: Executive of the Chamber of Commerce, the top official of the local labor council, the editor of the largest newspaper, the president of the local bar association, the president of the largest bank, the two chairmen of the local political parties, and the Mayor.

The participation of each city's citizenry in the decision-making process was measured by three questions administered to each of the cities. The items asked how many public meetings are held to discuss a typical city decision, how many persons come to such meetings and whether decisions are changed because of citizen response.

Does the index show enough reliability to be treated as a measure of anything? The sets of responses are so highly intercorrelated that a sophisticated test is not necessary. Each of the eight sets of responses was correlated against the mean score of all eight for that city. Since each contributes towards this mean there is, ofcourse, a spurious association. If the responses were truly uncorrelated, then each response would correlate with the grand mean with a gamma of about .25. The actual gammas are considerably higher than that: they average .67 and range from .41 (labor leaders) to .88 (Chamber executives). It is clear that the eight respondents are agreeing with each other in describing the degree of citizen participation in their city.

IV: Measuring the Power of Business

The power of the business sector is measured by a reputational question. Each of the same eight respondents was asked to consider five issue areas:- the selection of candidates for the school board, the passage of a municipal bond referendum, the adoption of an urban renewal project, the adoption of a program of air pollution control, and the selection of a mayor, and asked to consider whether the support of any of fifteen different groups should be considered essential, important, or not important in the selection of the candidate or the adoption of the program in question. The fifteen include five business groups: the Chamber of Commerce, retail merchants, industrialists, bankers, and a category called "other businessmen". The responses to these five groups by these eight respondents were averaged for the city to produce an overall score.

Again we must consider whether there is agreement among the eight respondents about the amount of influence that groups have. Using the average percentage for that group and issue across all cities, we can compute the expected number of times each particular group will be considered essential on a particular issue by all but one of the respondents in any particular city. The expected number of times that this should occur for each group and issue is totalled across the 75 combinations of groups and issues. When this is done we find that we can expect all but one of the respondents to agree that a certain group is essential only approximately 32 times. In fact, this happens 81 times. The differences between the extreme cities are impressive. In Albany, New York, for example, the Democratic party is judged "essential" 28 times out of 30 ratings, while no other group receives this rating more than four times. In contrast, in Palo Alto, neighborhood groups are singled out as essential 15 times while neither political party is ever given this rating.

Since we wish to measure the influence of the business community relative to other groups, the total influence score of the five business groups was compared to the scores of the other ten groups in the city and the final business influence score used was the deviation from a regression line through the data. Thus we are measuring the relative influence of the business community in comparison to other groups in the city.

V: Measuring the Power of Political Parties

The measures of the strength of political parties is based upon the following conception of a "strong" local political party: A "strong" party is one which can limit the number of people who have the opportunity to be elected to public office. It should be composed of an elite group of people within the party who are set apart from the rank and file because they are most active and because they share among themselves the spoils of political office.

Four items were used to provide an overall portrait of the strength of political parties in each of the fifty-one cities.

First, the subjective reputation of that party which controls most of the offices was judged by the eight informants at the same time as they were rating the influence of business groups. Secondly, the influence of the dominant political party on elections for mayor was singled out and given a heavier weight. A third item used to separate strong from weak parties is whether it is able to limit its candidates to those who are loyal party members. The strong party is one whose candidates come up through the ranks as reported by the party chairman. The fourth measure is the availability of government offices - patronage - to party officials. Presumably the party which has jobs can buy discipline with those jobs and can convert the loyal work of patronage employees into the ability to control a larger section of the electorate.

The four measures are highly intercorrelated. The Patronage-base parties recruit their candidates from the ranks of party regulars, as we all have assumed; the gamma is .87. When these two measures are combined into a single code they correlate with a gamma of .48 against the informant's rating of the party's influence on mayors elections and .40 with its influence on all five issues. The combined rating would be quite satisfactory except for the fact that there is a significant amount of missing data because of respondent refusal to submit to interviews or to answer certain questions, making it necessary to adjust for missing data.

VI: The Balance of Power Between Sectors

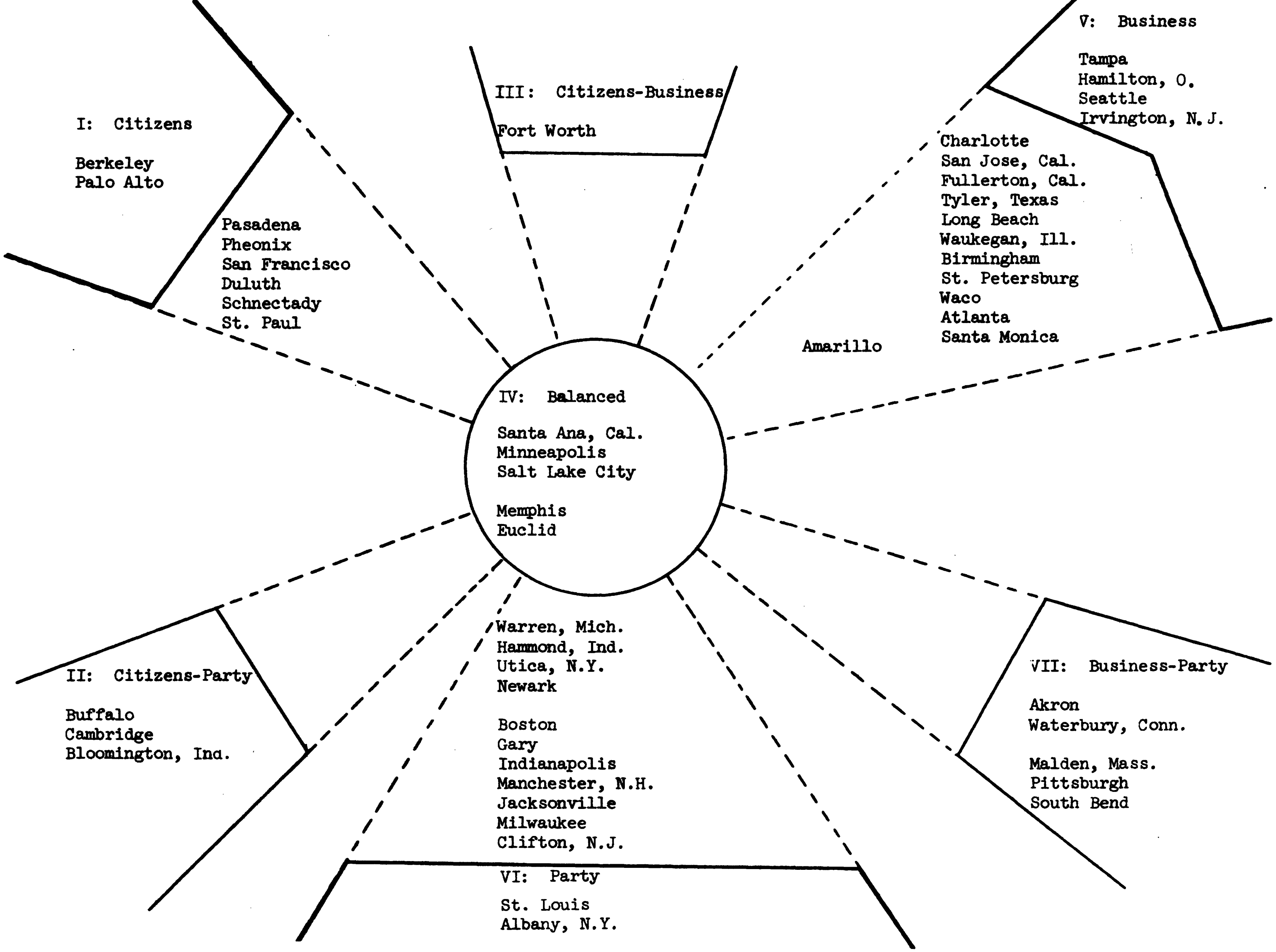
As we stated, we do not believe it possible to decide which sector has more power in absolute terms. For all we know, every city may be "really" ruled by business, or none may be. In a first look at the data, we assumed the three groups to be, on the average across all cities, roughly equal in power - the data are normalized.

The cities were divided into trichotomies on the three different measures of influence and the city was assigned into a cell in an overall typology depending on which of the three groups: business, local parties, or citizens, was most influential, or assigned a coalition cell in the table if two or all three of the groups were tied in influence. After this was done the typology seem somewhat unsatisfactory; For example Memphis, Tennessee, was described as a citizen-dominated city, which seems rather unlikely. The typology was then modified by scoring cities as either high or low on citizen participation, with the majority of the cities being dropped into the low category. One other change was made; party data were missing for Newark, N. J. and since we had a fairly good knowledge of the structure of the political parties there we supplied what we believed to be the correct answers to the questions which were not answered there.

Figure One shows the cities in the seven cells of the typology. When we compare the number of cases in each category with the expected frequencies derived from a random model, we find that each of the three pure types - the cities dominated entirely by citizens, entirely by business, and entirely by political parties - are more frequent than expected. There are 36 such cities, compared to an expectation of only thirty. The most frequent type is the business-dominated city. Of course, this distribution should not be taken too seriously since it is in part an artifact of the normalizing process described earlier.

A number of the business-dominated com-

FIGURE I: BALANCE OF POWER IN THE 51 CITIES



munities are southern cities and they are generally somewhat smaller, on the average, than cities in other categories. Political parties are frequently very weak in these cities and a number of them have city manager government. They are not, however, all clean government cities, nor are they necessarily all well administered cities. Conversely, the political party-dominated cities are not all dens of sin, although there are a few that might fit that description.

The most interesting categories are the citizen-dominated cities and the cities where the citizens and the political parties share power. The first case includes a number of "ideal" good government, weak party, mass movement cities of which Berkeley and Palo Alto, which score at one extreme on this citizen participation scale, are good examples, but also include Duluth, Schenectady and St. Paul. The latter may be in the sample because of the weakness of political parties and the strength of such citizen groups as ethnic groups and labor in these cities. In working-class cities such groups are the functional equivalent of citizen organizations. The citizen-party coalition is an unlikely coalition and indeed the three cities in that category are unlikely cities. Cambridge and Bloomington, Indiana are cities with traditional political parties but which live in the presence of articulate universities; it would be fair to talk about those two cities as having a citizen-party conflict rather than a citizen-party coalition. The other city, Buffalo has very weak political parties which are torn by ethnic conflict.

The Lincoln Steffins cities -- those where business and party leaders share the power -- are not common; there is a definite negative association between the amount of power these two groups have, $\gamma = -.46$, and there are only five such cities in the sample.

The residual category, which we have called balanced cities, are simply cities in which all three sectors are rated as having much influence, as in the case of Minneapolis, Santa Ana, or Salt Lake City, or alternatively where these sectors are seen as having very little influence, such as Memphis or Euclid.

VII: How the Balance of Power Affects Decision-Making

We next need to verify that this typology of cities is of some value. Intuitively, it conforms to the conventional wisdom that we have about such cities as Atlanta or Pittsburg but beyond that, it would be nice to know that we have in fact described the cities in such a way that we can predict how they will behave.

While we plan to carry on such an analysis, we are not optimistic; for while the typology tells us which group in the community dominates, it doesn't tell us what the group's political ideology is, or how efficient they are in obtaining their goals. In Berkeley, citizen participation reflects a very liberal ideology; in Pasadena it manifests itself in right wing radicalism. Atlanta, with its liberal business community, and Birmingham, with its conservative leadership, fall in the same category; or to look at it a different way, the strong political parties of Milwaukee are reasonably efficient in governing their city, while those in Jacksonville seem quite incapable of efficient government. Thus, we cannot expect very much clear difference in either the type of programs which the cities pursue or in the success in reaching their goals.

However, we can talk about the way in which the balance of power in each community controls, if not the outcomes of issues, at least the kinds of issues which come up. Bachrach and Baratz[5] in particular stress non-appearance of an issue as being of central importance in the study of power, for the power of a group may be reflected more in its ability to force an issue into the decision-making arena, or to prevent it from being brought up.

The eight respondents in our sample were asked to describe the major problems facing their city. These issues were then divided into four categories: first, issues of economic development and taxes; second, issues of providing services to the citizens; third, governmental reform; and fourth, issues of an ideological nature, either race relations or the whole issue of the amount of community conflict itself. When we look at the seven types of cities in Table 1, we see a fairly simple story. The cities in which business is either the controlling group or shares in the power are most likely to be concerned with economic development. Those cities in which the parties are strong or which are balanced are most concerned with services (or parenthetically, with government reform, which is not a separate category here); and citizen-dominated cities are most concerned with ideological issues.

There are only six cities in the sample which consider race relations their most serious problem. These data were gathered a couple of years ago, and the number may be higher now. Two of these, Pasadena and San Francisco, one would not think of as having especially severe racial tensions; but these are cities in which a high level of citizen participation has permitted a popular issue with mass appeal to rise to prominence. The other cities are party or business dominated. One, Atlanta, has worked very hard to avoid anticipated racial difficulty and the other three, Birmingham, Gary, and Waukegan, Illinois, have probably had the issue raised in a revolt against an inflexible power structure which suppressed the issue until it boiled over. In general, the issues

TABLE I: POWER TYPOLOGY AND MOST SERIOUS PROBLEM
AS SEEN BY INFORMANTS

Type of Problem	Citizens	Cit-Bus	Business	Balanced	Bus-Par	Party	Cit-Par
	TYPOLOGY: DOMINANT POWER SECTOR						
Economic Growth, Business Renewal, Tax base	4	1	9	1	4	2	1
Services to Citizens	0	0	3	3	1	7	2
Government Reform	0	0	0	1	0	2	0
Ideological Issues, Conflict	4	0	3	0	0	2	0
Total	8	1	15	5	5	13	3

Note: one city not coded

Summary: Business-dominated cities are more likely to say "economic dev": $\gamma = .56$
Party cities are more likely to say "services, reform": $\gamma = .61$
Citizen cities are more likely to say "ideology": $\gamma = .58$

which dominate the party-cities are often of this same rebellious character, and seem to have been raised against the wishes of the rulers. For example, all four of the cities which consider education to be their most serious problem are of this type.

With the small number of cases at our disposal, it is difficult to make a convincing argument. To strengthen our case, let us single out the eight most ideal cases, which are shown in the heavy boxes of the sundial in Figure 1. In both Berkeley and Palo Alto, the major issue is that there is too much conflict and mistrust. In the four business-dominated cities, the issues are taxes, the rate of population growth, and in two cases, economic growth. In the two extreme party-dominated cities, St. Louis and Albany, the issues are the tax base and housing. Thus, seven of the eight cities fit our conception very well. Citizen cities fight about things citizens are interested in; business cities worry about business; and party cities get screamed at for doing things wrong.

Next, let us ask how an issue is handled once it enters the arena in these cities. Table 2 looks at the seven types of cities and asks, compared to other cities of the same size, is there more or less conflict in each type of city? The controversy measure is based on the responses of our informants about the level of controversy in ten different issue areas, and the data has been standardized to remove the effect of city size, since large cities have more controversy than smaller ones. The table indicates that the highest level of controversy appears in three types of cities; one type, of course, is the citizen-dominated city; the other types are those where the parties are coequal in influence with either citizens or business. The findings are not statistically reliable, (whatever that means with this type of sample) but they are plausible.

TABLE 2: POWER TYPOLOGY AND LEVEL OF CONTROVERSY,
STANDARDIZED TO REMOVE THE EFFECT OF CITY SIZE

Level of Controversy	Citizen	Cit-Bus	Business	Balanced	Bus-Par	Party	Cit-Par
	TYPOLOGY: DOMINANT POWER SECTOR						
% High	75%	0%	38%	40%	80%	31%	67%
N	(8)	(1)	(16)	(5)	(5)	(13)	(3)

In a previous paper on the amount of controversy which is found in citizen-dominated cities, we suggest that the controversies are the result of the ease with which people can be mobilized to take part in debate. Cities in which political parties are strong but other groups are strong also may be conflict-prone because of the conflicting interests of the groups. This is certainly true when one thinks of the town-gown controversies in Cambridge or Bloomington, and might fit Buffalo, where the ethnic groups are pursuing goals at variance with the political parties which keep attempting to build coalitions between them. Similarly, we would guess that the five business-party cities are troubled by the traditional reform issues which have made these two groups rival over the years, and which results, in this sample, in a general negative correlation between the amount of influence each has; cities are either reformed, or they are not, and these cities may be struggling because they have not yet gone one way or the other. Controversy, of course, does not mean bad government: experienced observors would probably say that some very well-governed cities and some very badly-governed cities have high controversy.

VIII: Why Do Cities Have Different Power Balances?

If we have given the reader enough data to guess that our typology is valid, we will look quickly at the question of why different cities have different forms of power structures. In Tables 3 and 4 we look at only two variables, and we discover a fairly straightforward story there.

TABLE 3: PER CENT OF POPULATION COLLEGE GRADUATES, AND POWER TYPOLOGY

Typology: Dominant Power Sector	% College Graduates		
	0-6%	7-9%	10+%
Citizen	0	4	4
Cit-Bus	0	0	1
Business	4	4	8
Balanced	0	3	2
Bus-Par	3	2	0
Party	9	4	0
Cit-Par	1	0	2
Total	17	17	17

SUMMARY:

	0-6%	7-9%	10+%	
% of cities business-dominated	41%	35%	53%	(γ = .16)
% of cities citizen-dominated	6%	24%	41%	(γ = .62)
% of cities party-dominated	77%	35%	12%	(γ = -.77)

TABLE 4: SUBURBANIZATION OF MANAGERS, OWNERS, PROPRIETORS, AND POWER TYPOLOGY (CITIES OF OVER 100,000 POP.)

Typology: Dominant Power Sector	Ratio: $\frac{\text{\% Managers, Owners, Proprietors in City}}{\text{\% Managers, Owners, Proprietors in Metropolitan Area}}$		
	0 - .89	.90 - .99	1.00+
Citizens	3	0	3
Cit-Bus	0	1	0
Business	1	3	4
Balanced	1	3	0
Bus-Par	0	2	2
Party	7	3	0
Cit-Par	2	0	0
Total	14	12	9

SUMMARY:

	0 - .89	.90 - .99	1.00+	
% of cities business-dominated	7%	50%	67%	(γ = .75)
% of cities citizen-dominated	36%	8%	33%	(γ = -.14)
% of cities party-dominated	64%	42%	22%	(γ = -.53)

Table 3 indicates that citizens tend to be influential where they have the resources for leadership; citizen-dominant cities have large numbers of well-educated people; and these are the cities where parties have been weakened by reform. Second, we see in Table 4 that businessmen have influence where they have a stronger interest; namely, in cities where they have not yet fled to the suburbs.

Needless to say, all of this analysis is in a very preliminary stage, but we think it is clear that the ability to gather standardized data over a large enough number of cities to permit traditional survey-style statistical analysis with cross-tabulations will pay some dividends.

The data described here was gathered under a grant from the National Science Foundation to determine the feasibility of this approach, and the data will be used for a more extensive study of the issues of air-pollution and urban renewal. The Johns Hopkins University, the National Opinion Research Center, and a number of political scientists across the nation will be involved in creating what we have called a permanent community sample of 200 cities of over 50,000 population, including all cities of over 150,000 population in the United States.[6] Studies of school desegregation and of the local Community Action Agencies are underway in some of these cities, and other studies are being either designed, discussed, or dreamed about. Since the sample is indeed permanent, each study will contribute to a data archive which will eventually provide for all of us the opportunity to get at the questions we have all wondered about; what types of cities have riots, for example; or what difference does it make whether the local newspaper is the St. Louis Post Dispatch or the Chicago Tribune? Should Washington, D. C. have an elected School Board? Why can't Milwaukee have major league baseball and why does Pittsburg have a good symphony? And eventually it may be that we can consider again the question political scientists have given up hope of answering; what type of governing institutions should cities have to accomplish the things they want to do?

Footnotes:

1. Matthew Crensen contributed a number of valuable suggestions for the data analysis given here. James J. Vanecko directed the study from which the data are taken.

2. Robert A. Agger, Daniel Goldrich, and Bert E. Swanson, The Rulers and the Ruled, New York, Wiley, 1964.

3. Floyd Hunter, Community Power Structure, Chapel Hill, University of North Carolina Press, 1953.

4. Robert A. Dahl, Who Governs? New Haven, Yale University Press, 1961.

5. Peter Bachrach and Morton S. Baratz, "Two Faces of Power", American Political Science Review, 56, December 1962, pp. 947-952.

6. Peter H. Rossi and Robert Crain, "The NORC Permanent Community Sample," The Public Opinion Quarterly, Vol. 32, Summer 1968, pp. 261-272.

CONCEPTUALIZING AND MEASURING POLITICAL INVOLVEMENT OVER TIME: A STUDY OF BUFFALO'S URBAN POOR*

Everett Cataldo and Lyman Kellstedt
State University of New York at Buffalo

I

The "dilemma" in American race relations that Gunnar Myrdal once wrote about[1] has now reached crisis proportions. Rioting has become a means of expressing disapproval and frustration over the lack of responsiveness of American political and social institutions to meet satisfactorily the needs of the Black ghetto. Rioting as a form of political behavior has been the subject of much attention in private discussions, the media, and ultimately by a commission appointed by the President. The latter, the U.S. Riot Commission, attempted to explain why riots have occurred and what might be done to prevent their recurrence.[2]

Implicitly, at least, the effort to understand the riots of recent years raises the question of whether Black Americans have abandoned conventional political means (voting, working within political parties, and the like) for more unconventional means. Recent studies of urban disorders have tended to focus on unconventional political behavior, while the literature of political science traditionally has done just the opposite. Our concern, therefore, is to compare participation rates for Blacks and Whites, using both conventional and unconventional measures of political participation, and to isolate some of the important correlates of this participation. Broadly, the questions that concern us include:

1. What are the different rates of conventional and unconventional participation for Blacks and Whites?

2. What effect do social status variables have on political participation?

3. What effects do attitudinal variables have on participation?

4. How are conventional and unconventional participation related?

Before presenting an analysis of data bearing on these questions, it would be useful to place our study of participation within the body of literature concerned with political participation and to make clear some of our central assumptions.

Notions of participatory democracy are in a period of revival. Historically, these ideas were part of classical democratic theory, and we are reminded of the Populist movement of the late 19th and early 20th centuries with its concern for participatory democracy. The legislation that created the War on Poverty seemed to accept this idea. There was to be "maximum feasible participation" of the poor in the solution of their own problems. The assumption was made that government would be more democratic and useful if it were returned "to the people." Although some erosion of the principle of "maximum feasible participation" has taken place in the War on Poverty, there is no doubt that participatory democracy continues to be a potent organizing principle at this point in history.

Participatory democracy has been vigorously attacked from time to time by both scholars and social critics. Peter Bachrach labels one source of these attacks as "the theory of democratic elitism."[3] The assumption of democratic elitists is that the principle of participatory democracy is outmoded because the masses are basically apathetic about political matters. Elites become the forces in the society that sustain the system. Bachrach summarizes the democratic elitist view thusly:

> To be sure, the ordinary man still plays a role in the system since he has the freedom to vote, to bring pressure upon political elites, and to attempt himself to rise to an elite position. But by and large he does, and is expected to, remain relatively passive -- in fact the health of the system depends upon it.... Thus the political passivity of the great majority of the people is not regarded as an element of democratic malfunctioning, but on the contrary, as a necessary condition for allowing the creative functioning of the elite.[4]

Bernard Berelson, in a famous study of voting behavior, has emphasized another reason for the view that citizen apathy is functional for the system: "If everyone in the community were highly and continuously interested, the possibilities of compromise and gradual solution of political problems might well be lessened to the point of danger."[5] Thus, he continues, democracy requires" ... a body of moderately and discontinuously interested citizens within and across social classes, whose approval of or at least acquiescence in political policies must be secured."[6] System stability, therefore, is enhanced not by maximum participation but by moderate levels of participation.

Such a position directly confronts the position of classical democratic theory. John Stuart Mill argued that not to engage in political affairs would result in the stunting of man's intellectual and moral capacities.[7] As Bachrach argues, political involvement ".... is an essential means to the full development of individual capacities,"[8] and that " .... the majority of individuals stand to gain in self-esteem and growth toward a fuller affirmation of their personalities by participating more actively in meaningful community decisions."[9] We accept this argument and feel that it is particularly important for the Black American. Held back by an oppressive slave system and by legal and extra-legal discriminatory measures, the Black man has found it more difficult to develop a sense of self-esteem, or identity, if you will. Lacking family ties as close-knit as those of other ethnic minorities, which were so central to their developing a sense of identity, the Black man may need to participate in politics for purposes of self-development as well as for material payoffs.

In addition to missing the important point about the function of participation for individuals, the theory of democratic elitism fails to recognize the latent function that citizen participation performs for the system.[10] Conventional participation involves working within the system and thus builds support for the system. This is the case even if the activities are directed toward changing things; participation of the conventional sort is not directed toward overthrowing the system. Political participation thus performs two important functions that the democratic elitist theory ignores: 1) it functions to develop individual personalities more fully than non-participation; and 2) it functions as an important support mechanism for the system.

Much of the debate between democratic elitists and their critics may have been rendered academic by now. Whether one regards maximum citizen participation as functional or not, increased citizen participation can be noted especially among Blacks. Among other reasons, increased participation by Blacks can be accounted for by:

1) rising expectations which lead to increased participation;

2) the large number of self-help organizations that have risen in the ghetto; and

3) the activities of government.

Will participation by Blacks take the form increasingly of violent upheaval against the system, or will it involve working within the system to affect change? If the latter, we feel it will function both to assist the Black man to develop a sense of identity, and to buttress the democratic system.

Matthews and Prothro define political participation as "all behavior through which people directly express their political opinions."[11] In their study of Negro political participation in the South, Matthews and Prothro developed a political participation scale along four dimensions: talking politics; voting in elections; participating in campaigns; and holding office or belonging to a political group.[12] Milbrath has suggested a continuum of political activity that extends from low level "spectator" activity to higher level "gladitorial" activity. He lists fourteen activities ranging from "exposing oneself to political stimuli" at the lowest level, to "holding public and party office" at the highest level.[13] The list is hierarchical; that is, it extends from those behaviors most often performed to those least often performed. Thus political participation is said to be cumulative; individuals who engage in the "gladiatorial" activities are assumed to engage in the lower level activities as well, but not vice-versa. The striking characteristic of all these activities is that they are conventional in the sense of being considered socially legitimate and acceptable. Unconventional activities, such as public demonstrations and riots, are excluded from the hierarchy of political involvement.

Recent events indicate that political participation involves more than voting, working in political campaigns or even holding public office. In recent years, some of the most significant changes in public policy in the U.S. have followed in the wake of highly unconventional and disruptive political behavior by significant segments of the population. The compelling questions about political participation, therefore, are what kinds of activities are being engaged in by what individuals, for what purposes, and with what consequences for individuals and for the political system.

In attempting to answer these questions, we have cast a broad net in terms of the political activities engaged in by citizens. In 1966-67, a sample of over 1,000 citizens in the Buffalo area was interviewed. Follow-up interviews are currently in the field with the original sample and a control group. Both interviews have contained upwards of 30 participation items which range from keeping informed about politics and voting on the most conventional end, to joining in public street demonstrations and rioting on the most unconventional end. Rather than developing the items in hierarchical or cumulative fashion only, we have tried to devise analytically meaningful categories of participation each containing several items.

Five categories or dimensions of participation have been developed. These are: 1) Citizen Duty; 2) Support of Authorities; 3) Partisan Involvement; 4) Opinion Leadership; and 5) Protest-Demonstrate. The citizen duty dimension involves such basic activities as voting and demonstrating patriotism. The support of authorities dimension involves support for public officials when their authority or the law is challenged. Included in the partisan involvement dimension are the party and campaign activities common to measures of political participation in other studies. Opinion leadership involves, among other things, keeping informed about politics, making one's political views known, and attempting to influence others in their political preferences. The protest-demonstrate dimension consists of a set of items ranging from sending protest messages to public officials and attending protest meetings to demonstrating and rioting.

The development of these dimensions of participation permits an analysis of the kinds of generic activities performed by significant sub-groupings within the population. What groups are most supportive of legalized authority? Are the poor less supportive than higher status individuals? Are Blacks less supportive than Whites? Are activities on the unconventional protest-demonstrate dimension inversely related to the more conventional activities on the partisan involvement dimension? Do Blacks see standard forms of political participation as means by which they can fulfill their expectations, and do they engage in them to any significant degree? How are such attitudinal variables as sense of self-esteem, political efficacy and political cynicism related to conventional and unconventional forms of participation? Are those who feel less efficacious and more cynical disposed toward unconventional forms of participation? These are illustrative of the kinds of questions that can be explored with conceptually meaningful categories of political participation. Moreover, by repeating items from one survey to another variations in the kinds and rates of participation can be explored. Are increased levels of supportive and partisan activities associated with rising levels of income and education? Are increased levels of inefficacy and cynicism accompanied by an increased tendency to use unconventional means to achieve political goals and express political preferences? Will an increase over time in partisan involvement be accompanied by a decreased propensity to engage in activities of the protest-demonstrate variety?

II

We shall now address ourselves to some of these questions about political participation by presenting data from the survey of 1966-67. Many of the findings presented here derive from a card sorting technique.[14] Respondents were given cards with 21 participation items printed on them. They were asked to sort the cards as to the degree to which they performed the activities (1. Never; 2. Seldom; 3. Fairly Often; 4. Regularly) and the degree to which they felt a responsibility to perform them (1. None; 2. Some; 3. Important; 4. Essential responsibility).

One of our basic research questions was whether Whites and Blacks differed in their performance of conventional and unconventional political activities. Particularly, we were interested in discovering any tendency among Blacks to reject conventional modes of participation in favor of unconventional ones. Given the lower status of Blacks in the community[15] and the rising incidence of rioting in Black ghettoes over the last few years,[16] one might expect Blacks to participate in conventional activities at lower levels than Whites, and to show a greater propensity to engage in active demonstrations of protest. Our conventional participation measure consists of several items along the citizen duty, partisan involvement and opinion leadership dimensions. Table 1 lists the conventional participation items for Blacks and Whites with mean scores for each item. As can be seen, Blacks do not differ appreciably from Whites in their conventional participation. On the partisan involvement dimension and the citizen duty dimension there is a slight tendency for Blacks to score higher. On the opinion leadership dimension there is a slight tendency for Whites to score higher.

Our 1966-67 survey contained two items in the area of unconventional activity, mean scores for which are presented in Table 2. It is clear that neither racial group engages in street demonstrations or riots to any great degree. The scores, however, are higher for Blacks than for Whites. Since participation in these unconventional ways is rare, perhaps it is more meaningful to inquire into how much responsibility individuals feel they have to engage in these activities. The scores are again low, but they are higher for Blacks.

While these data show a higher riot propensity among Blacks than Whites, they show as well that Blacks engage at least as much as Whites in conventional forms of political participation. The data would seem to suggest, then, that as a group Black citizens in Buffalo are not rejecting conventional modes of participation for unconventional ones. Rather, it would appear that Blacks are more highly politicized generally. It might be speculated, of course, that Blacks participated even more highly in conventional ways before our study began, and that their conventional participation, therefore, is decreasing. Although we do not have survey data on these matters prior to 1966-67, such a conclusion nevertheless is difficult to accept. In fact, what evidence we have points in the other direction. Buffalo's Black community seems to have become increasingly politicized over the past several years. Steady increases have taken place in voter registrations, voting and the formation of political organizations in the east-side ghetto. We would expect data from our current survey to show further increases in these activities whether this was a Presidential year or not.

These participation data suggests that Blacks are at least somewhat confident that their goals can be achieved through conventional means. If it could be shown, however, that those Blacks who are riot-prone (feel at least "some responsibility" to riot) participate in the more legitimate activities to a significantly lesser extent than those Blacks who are not riot-prone, then there would be some evidence that the more militant have lost confidence in their ability to accomplish ends through conventional means. As Table 3 shows, however, riot-prone and non-riot-prone Blacks are quite similar in their conventional participation activities. Small percentage differences exist between the groups on certain items, the riot-prone being higher on some, the non-riot-prone higher on others. These differences, however, tend to cancel each other out. The two groups simply are not appreciably different.

TABLE 1

CONVENTIONAL PARTICIPATION, BY RACE

Dimensions	Blacks	Whites
Citizen Duty		
*1. Registered to vote	3.45	3.37
2. Vote in elections	3.66	3.48
Partisan Involvement		
1. Join and support a political party	2.32	2.16
2. Take an active part in a political campaign	1.68	1.57
3. Participate in a political party between elections	1.72	1.68
Opinion Leadership		
1. Keep informed about politics	2.73	3.00
2. Engage in political discussion	2.34	2.39
3. Discuss politics between elections	2.13	2.23
4. Inform others about politics	2.10	1.91
*5. Tried to influence political decisions other than by voting	1.44	1.56
*6. Have been asked for advice about politics	1.53	1.46

*These items did not appear on the card-sort, but elsewhere on the interview schedule. The response categories for them were "yes" or "no." Mean scores for these items were made somewhat comparable to those on the four-point scale by scoring a "4" for "yes" and a "1" for "no."

TABLE 2

UNCONVENTIONAL PARTICIPATION ITEMS,* BY RACE

	Blacks		Whites	
	Performance	Responsibility to perform	Performance	Responsibility to perform
1. Join in public street demonstrations	1.39	1.56	1.11	1.20
2. Riot if necessary to get public officials to correct political wrongs	1.23	1.43	1.10	1.18

*These items correlate highly with each other (.41 for Whites and .46 for Blacks) but not with any of the other participation items.

Thus even among the more militant Blacks there appears to be no tendency to eschew conventional modes of participation for unconventional modes.

The data presented thus far reveal an important aspect of the relationship between Black citizens and their political communities that the recent emphasis on urban violence may be obscuring -- that is the degree of support for the political system manifested by the rates of conventional participation we have examined here. The reports of the U.S. Riot Commission (including the Supplemental Studies)[17] have emphasized riot behavior among Blacks and

TABLE 3*

PARTICIPATION IN CONVENTIONAL POLITICAL ACTS FOR RIOT-PRONE AND NON-RIOT-PRONE BLACKS

Dimensions	Non-Riot-Prone: Not at all-Seldom (No)	Non-Riot-Prone: Fairly often-Regularly (Yes)	Riot-Prone: Not at all-Seldom (No)	Riot-Prone: Fairly often-Regularly (Yes)
Citizen Duty				
**1. Registered to vote	20%	80%	12%	88%
2. Vote in elections	10	90	10	90
Partisan Involvement				
1. Join and support a political party	56	44	58	42
2. Take an active part in a political campaign	83	17	75	25
3. Participate in a political party between elections	81	19	82	18
Opinion Leadership				
1. Keep informed about politics	37	64	44	56
2. Engage in political discussion	60	40	53	47
3. Discuss politics between elections	72	28	73	27
4. Inform others about politics	66	34	67	33
5. Tried to influence political decisions other than by voting	83	17	91	9
6. Have been asked for advice about politics	72	28	73	27

*The N for the non-riot-prone is 199; for the riot-prone it is 67.

**The registration item asked specifically if respondents were registered to vote in the election of November, 1966. The voting item appeared on a card sort and inquired into the frequency of voting generally. Some respondents who reported that they voted regularly may not have been registered to vote in that particular election; for example, newcomers who had not satisfied residency requirements. We would not necessarily expect, therefore, means and percents for voting to be only equal to or lower than those for registration.

the disruptive consequences for the system which flow from this behavior. For all the value the reports may have in arousing the conscience of White America toward injustice, their concentration on unconventional behavior has given us an incomplete and perhaps misleading picture of the consequences of political activism by Blacks. Indeed, the findings of the reports may lead some to draw the wrong conclusions about the loyalty of Black Americans. An examination of the full range of political participation by both Blacks and Whites gives a more balanced view. We do not see here evidence of a Black community refusing to employ the traditionally legitimate and acceptable ways of expressing political preferences. To the contrary, what we see in these data is a Black community apparently willing to work through the system and with it in the same ways that White Americans always have, even though Blacks exhibit a somewhat greater propensity to engage in unconventional behavior. Acceptance of the norms of conventional political participation has important implications for system support among Blacks that should not be overlooked. Especially dramatic in this respect is the finding that conventional behavior is apparently the normal modus operandi even of those who, under certain circumstances, countenance breaking the law and engaging in disruptive behavior to achieve their goals.

III

To understand in more detail what leads individuals to participate in politics in different ways and at different rates, we move now to a discussion of the correlates of political participation. Specifically we shall explore the relationships between participation and social status, political cynicism, political efficacy, political information, and sense of progress and optimism for the future with respect to the self and one's political communities.

Numerous studies have shown that political participation increases with increased social status.[18] Table 4 displays product-moment correlations for Blacks and Whites between education and income and the measures for conventional and unconventional participation. For both Blacks and Whites there is a tendency for participation in conventional ways to increase with increased education and income, although some of the correlations are relatively weak. For both racial groups status correlates most highly with talking about politics. This form of participation involves the use of verbal skills; and, generally speaking, the higher one's education the more likely he is to be skilled verbally and to enjoy political discourse. To engage in this activity to any great extent, it may also be necessary to have sufficient leisure time or the right kind of work situation. The higher one's income the more leisure time he may have, and the more likely he is to have the kind of job that permits conversation and the opportunity to exchange ideas. The next highest correlations are found between status and informing others about politics, trying to influence political decisions other than by voting, and being asked advice about politics. Again, verbal skills, time and some degree of prestige among one's peers are involved in performing these activities.

Except for the higher positive correlation between income and demonstrating for Blacks, there is little association between status and propensity toward unconventional acts. The .22 correlation between income and demonstrating suggests that among some middle or upper income Blacks, demonstrating may not be viewed as unconventional or illegitimate. It may, indeed, be viewed as occasionally necessary to sustain gains or to pave the way for new ones. One might expect the correlations between both income and education with the demonstration item

TABLE 4

SOCIAL STATUS AND POLITICAL PARTICIPATION, CORRELATED BY RACE

Dimension	Blacks Education	Blacks Income	Whites Education	Whites Income
Citizen Duty				
1. Registered to vote	.14	.12	.14	.19
2. Vote in elections	.11	.14	.13	.21
Partisan Involvement				
1. Join and support a political party	.19	.09	.19	.18
2. Take an active part in a political campaign	.16	.20	.21	.19
3. Participate in a political party between elections	.17	.19	.10	.15
Opinion Leadership				
1. Keep informed about politics	.28	.08	.27	.17
2. Engage in political discussion	.32	.21	.34	.32
3. Discuss politics between elections	.31	.27	.35	.26
4. Inform others about politics	.22	.18	.25	.22
5. Tried to influence political decisions other than by voting	.24	.28	.31	.27
6. Have been asked for advice about politics	.20	.18	.20	.24
Protest-Demonstrate				
1. Join in public street demonstrations	.06	.22	.05	.04
2. Riot if necessary to get political wrongs corrected	-.04	-.02	-.07	-.06

to be more nearly identical among Whites than among Blacks given the closer association for Whites between income and education.[19]

The findings of other studies that political participation increases as status increases are confirmed here, but only with respect to conventional political activities, and then not in a particularly pronounced way. When a propensity to engage in unconventional participation is measured a different pattern emerges. The effects of status generally are greatest on opinion leadership activities which require verbal skills, time and some degree of prestige. The other conventional activities seem to require no special skills and the masses are encouraged to perform them. The relationships of these items with status are, therefore, correspondingly lower. The fact that partisan involvement and citizen duty activities are frequently engaged in by lower status individuals has implications for system stability. Widespread participation in these activities by both Blacks and Whites strengthens such existing institutional patterns as political parties and the electoral system and thus helps maintain system equilibrium.

Political participation has been shown to increase with increased political efficacy and with decreased political cynicism.[20] Our respondents were scored on each of these variables on the basis of responses to several items in each category, with the higher the score, the higher the possession of the attribute. The efficacy category consists of items dealing with an individual's confidence in his ability to effect governmental decision-making, and a belief that government is responsive to citizens. The cynicism items deal with an individual's suspicion of the motives and behavior of public officials.

Table 5 reports correlations for Blacks and Whites between political participation and cynicism and efficacy. As can be seen, the inverse relationship between cynicism and conventional participation is consistent for both racial groups; as cynicism increases, conventional participation tends to decrease. Although the magnitudes of the individual correlations are not great for either group, the observed relationship, for the most part, is slightly more pronounced for Whites than for Blacks. The relationships between efficacy and acts of conventional participation are also

TABLE 5

POLITICAL PARTICIPATION, CYNICISM AND EFFICACY, BY RACE

Dimension	Blacks		Whites	
	Cynicism	Efficacy	Cynicism	Efficacy
Citizen Duty				
1. Registered to vote	-.13	.15	-.03	.16
2. Vote in elections	-.09	.16	-.07	.18
Partisan Involvement				
1. Join and support a political party	-.04	.25	-.16	.19
2. Take an active part in a political campaign	-.04	.10	-.12	.15
3. Participate in a political party between elections	-.06	.14	-.09	.10
Opinion Leadership				
1. Keep informed about politics	-.06	.25	-.12	.30
2. Engage in political discussion	-.08	.23	-.17	.27
3. Discuss politics between elections	-.07	.22	-.15	.24
4. Inform others about politics	-.09	.15	-.10	.19
5. Tried to influence political decisions other than by voting	-.10	.26	-.09	.22
6. Have been asked for advice about politics	-.12	.24	-.07	.14
Protest-Demonstrate				
1. Join in public street demonstrations	.03	-.04	.02	.03
2. Riot if necessary to get political wrongs corrected	.18	-.17	.12	-.01

consistent for each racial group; as efficacy increases, there is a tendency for conventional participation to increase. The correlations are of somewhat greater magnitude than those between cynicism and participation. Just as with social status, the effects of efficacy are strongest on opinion leadership activities. To engage in these activities may require a greater amount of confidence than is needed to vote or to join a party. Talking about politics, giving others political advice or trying to be influential are less likely to be attempted by individuals lacking confidence in their ability to communicate political ideas effectively or to be persuasive.

Neither cynicism or efficacy correlate significantly with acts of demonstration. The correlations between propensity to riot and cynicism are positive for both racial groups. For Blacks, efficacy and propensity to riot are inversely related; oddly enough, there is virtually no association for Whites.

Why might the association between cynicism and conventional participation be as low as it is for Blacks? Our survey data show Blacks to be more cynical about politics than Whites.[21] Years of discrimination and exclusion from meaningful participation in public affairs were highly likely to lead to cynical attitudes. In responding to interview items it is perhaps automatic for Blacks to give "cynical" answers. On the other hand, much that has gone on in recent years may have given Blacks hope that things are finally changing. Blacks are more prominent in public life than at any other time in our history; three major American cities have Black mayors; many of the more significant programs developed in the U.S. in recent years have been aimed at improving the lot of the Black man; and, to put the local situation in proper perspective, Buffalo's present mayor, though a White ethnic, has been identified more closely with the aspirations of Buffalo's Black citizens than any previous mayor of the city. These kinds of developments may explain, at least partially, why our 1966-67 survey showed Blacks to be optimistic about their own personal development and about the future of the U.S. and the Buffalo community. It may be that this sense of optimism is helping Black people to overcome the dampening effects on participation that cynicism normally breeds, even though Blacks themselves may feel that they have had good reason up to now to be cynical about politics. In any event, the relatively weak effect that cynicism has on the acceptance by Blacks of socially approved modes of political behavior might be viewed as another indicator of strong support for the political system by Blacks.

Matthews and Prothro found an individual's level of political information to be highly predictive of his level of participation.[22] Our respondents were given political information scores based on their ability to identify selected public leaders. As can be seen from Table 6, in most instances political information correlates more highly with conventional participation for both Whites and Blacks than did education, income or political efficacy. Status and efficacy were more strongly related to opinion leadership activities than to any of the other participation dimensions. Here we note that the highest correlations are on some of the opinion leadership items. At the same time, however, information is more highly correlated with some of the campaign activities and citizen duty items than with some opinion leadership items. Apparently the effects of information are more generally distributed among the various conventional participation dimensions than status or efficacy.

When we look at unconventional participation, however, the pattern is different. For Whites there is a very slight inverse relationship between information and unconventional participation, but this is not so for Blacks. In neither case does it appear that knowledge about an individual's level of political information will be a good predictor of that person's unconventional political behavior.

One might assume that levels of participation would be related to a personal sense of progress and a sense of optimism for the future and to a sense of progress and a sense of optimism for the nation and the community. The more progress and optimism an individual felt in these respects, the more likely he would be to participate. On the other hand, those who had little sense of progress and not much hope for the future might be expected to participate in conventional activities at much lower levels, and perhaps be more prone toward unconventional activity as a manifestation of frustration. This model assumes a kind of causal path from a sense of progress and a sense of optimism to high levels of conventional participation and low levels of unconventional participation. The model tends to assume that an individual rationally perceives his sense of progress and optimism for the future and links these conceptions with the different types of participation. Were the model found to be valid empirically it might bode ill for the system, for individuals who sense little or no progress (or who even sense retrogression) and are not optimistic either about their own future or the future of the communities to which they are tied might be prone to engage in behavior disruptive for the system.

On the other hand one may choose to emphasize the notion that participation takes place irrespective of where the individual sees himself in terms of a sense of progress and a sense of optimism and irrespective of whether he sees the national and local communities as progressing and views their future optimistically. This model would tend to see participation as flowing more from a sense of civic duty than from a rational conception of one's own interest. This model would, in addition, emphasize the strong system supportive function of participation.

To measure an individual's sense of progress and optimism with respect to himself, the

TABLE 6

POLITICAL PARTICIPATION AND POLITICAL INFORMATION

Dimensions	Blacks	Whites
Citizen Duty		
1. Registered to vote	.29	.22
2. Vote in elections	.22	.29
Partisan Involvement		
1. Join and support a political party	.29	.28
2. Take an active part in a political campaign	.29	.27
3. Participate in a political party between elections	.29	.18
Opinion Leadership		
1. Keep informed about politics	.32	.35
2. Engage in political discussion	.32	.33
3. Discuss politics between elections	.32	.34
4. Inform others about politics	.27	.32
5. Tried to influence political decisions other than by voting	.22	.24
6. Have been asked for advice about politics	.10	.20
Protest-Demonstrate		
1. Join in public street demonstrations	.08	-.04
2. Riot if necessary to get political wrongs corrected	.01	-.06

U.S. and the Buffalo community we used the "Self-Anchoring Striving Scale" developed by Hadley Cantril.[23] Respondents were asked to place themselves, the U.S. and Buffalo on an eleven point scale represented by a ladder with rungs ranging from 0-10, the bottom of the ladder signifying things at their worst, the top of the ladder indicating things at their best. In each case placements were made for the present, five years in the past and five years in the future. By subtracting past from present placements we devised a "Sense of Progress" index. Subtracting present placements from future placements gave us an "Optimism for the Future" index. On each of these respondents could rate themselves or their political communities as worse off, the same, or better off. Scores on these indexes were then correlated with participation scores. For Whites the correlations between conventional participation items and a sense of progress and sense of optimism for the self, the U.S., and Buffalo are negligible. In fact, 50 of the 66 possible correlations range from +.05 to -.05. The correlations for Whites on the unconventional participation items are low as well; only one of the 12 possible correlations lies outside the range of +.05 to -.05. Apparently, for Whites, decisions about engaging in political activities are made quite apart from any explicit considerations about a sense of progress and a sense of optimism.

Optimism for the futures of the U.S. and Buffalo is somewhat more positively related to the participation rates of Blacks. Earlier in this paper it was suggested that events of recent years may have given Blacks a sense of hope for the future, and that this was having a bearing on their conventional participation. When comparing Blacks with Whites the relationship between participation and optimism is more pronounced for Blacks on such activities as keeping informed about politics, informing others about politics, engaging in political discussion, voting in elections, taking an active part in a political campaign, and participating in a political party between elections. These findings, however, must be interpreted very cautiously. The argument that a sense of optimism influences participation rates implies a direction of causality that the correlation coefficients, by themselves, do not prove. It is entirely possible that for some increased participation has the effect of developing within them a sense of optimism. Even more important, perhaps, is the fact that even the strongest correlations between optimism and participation for Blacks explain very little of the total variance in their participation rates, the highest correlations being only .18. There is no evidence to support the hypothesis that unconventional acts will be engaged in more by those Blacks with little sense of progress or optimism. If anything the reverse is the case, with a sense of personal progress being positively related to a propensity to engage in these acts. (The correlations are

.26 for the demonstration item and .11 for the riot item.)

These data lead one to conclude that for Whites basic decisions about whether to engage in political activities are made quite apart from any explicit considerations about progress and prospects for the future; as suggested above, participation may spring from a sense of civic duty. In addition, a desire to conform to what is thought to be socially acceptable behavior, or simply habit may also be operative. For Blacks a slightly different pattern emerges. There is some tendency for Blacks who are optimistic about the future of the U.S. and Buffalo communities to participate more extensively in a conventional manner. Possibly, the use of politics to advance group interests is at work here.

IV

Our final task is to examine the relationships between the participatory activities presented in this paper. Table 7 shows the intercorrelations of each activity with each other activity for both Whites and Blacks. Certain clustering patterns can be observed for both racial groups. Opinion leadership activities -- keeping informed, discussing politics, etc. -- are highly intercorrelated. Partisan involvement activities -- joining and supporting a party, participating in it regularly, and working in a political campaign -- are also highly intercorrelated. In addition, certain items in each of these two categories correlate highly with items in the other. For Whites, informing others about politics has correlations of .51 and .46 respectively with taking an active part in a political campaign and participating in a political party between elections. For Blacks, informing others has correlations of .50 and .52 with those items. The rest of the items in the opinion leadership category correlate at moderate levels with the rest of the items in the partisan involvement category. The citizen duty items (registering and voting) correlate highly only with each other. We would expect this, knowing that many individuals do not participate beyond registering to vote and voting.

For both racial groups the two unconventional acts correlate reasonably highly with each other, but not with any of the conventional activities. For Whites all the correlations between propensity to riot and conventional activities are at or near zero, except for the citizen duty items. Registering to vote and voting correlate at -.14 and -.10 respectively with propensity to riot. For Blacks citizen duty items are not negatively correlated with rioting. A slight tendency was noted above for Blacks to participate in conventional ways for more purposive reasons than Whites. Perhaps for some Blacks, unconventional behavior is also viewed as purposive. For Whites such does not appear to be the case, and this may be especially so for Whites who do not participate beyond voting. For them voting may be done out of a sense of citizen duty only, and they would be strongly opposed to engaging in activity that would violate good citizenship norms.

These data suggest that conventional and unconventional forms of participation are not seen as diametrically opposed and mutually exclusive kinds of activities, for if they were, the demonstrate and riot items would have strong negative correlations with the conventional participation items.

V

This paper has compared Black and White citizens of an urban area on measures of conventional and unconventional political participation. One of its central questions was whether a strong tendency existed among Blacks to substitute unconventional behaviors for conventional ones in an effort to achieve political goals or to demonstrate hostility toward the system. No such tendency was observed. While a greater propensity to demonstrate and riot was observed among Blacks, they were seen to participate in conventional ways equally as much if not more so than Whites. Even those Blacks who show a propensity to riot participate in conventional ways as much as Blacks who are not riot-prone.

The factors that help explain differential rates of participation for Whites generally do so for Blacks as well. For both racial groups, conventional participation tends to increase with an increase in social status, political efficacy and political information, and to decrease with an increase in political cynicism, although this latter relationship is somewhat stronger for Whites than for Blacks. A sense of progress and optimism for the future has no relation to participation rates for Whites, but for Blacks there is a very slight relationship between optimism for the U.S. and Buffalo and conventional participation rates. This suggests, perhaps, the beginnings of a linkage for Blacks between decisions to participate and their future hopes, making participation for them a purposive activity. The low correlations between optimism and participation, however, justify advancing this only tentatively.

A propensity to engage in unconventional activities was seen to be essentially unrelated to social status and information level. For both Whites and Blacks riot-proneness increases slightly as cynicism increases. For Blacks riot-proneness decreases slightly as efficacy increases. Contrary to what might be expected, propensity toward unconventional activity is not related for either racial group with a sense of despair or lack of optimism for the future.

Finally, and perhaps most importantly, we do not find conventional behavior to be inversely related to unconventional behavior. The correlations between propensity to demonstrate or riot and conventional participation items are for the most part very weak.

TABLE 7*

INTERCORRELATIONS BETWEEN PARTICIPATORY ACTIVITIES

	Registered to vote	Vote in elections	Join and support a political party	Take an active part in a political campaign	Participate in a political party between elections	Keep informed about politics BLACKS	Engage in political discussion	Discuss politics between elections	Inform others about politics	Tried to influence political decisions other than by voting	Have been asked for advice about politics	Join in public street demonstrations	Riot if necessary to get political wrongs corrected
Registered to vote		.64	.32	.27	.17	.18	.12	.10	.20	.17	.15	.13	.06
Vote in elections	.65		.27	.16	.16	.31	.11	.11	.19	.09	.04	.10	.06
Join and support a political party	.25	.30		.44	.51	.39	.28	.14	.41	.22	.15	.08	-.01
Take an active part in a political campaign	.17	.18	.48		.59	.26	.37	.29	.50	.23	.18	.22	.17
Participate in a political party between elections	.20	.21	.49	.62		.37	.37	.35	.52	.25	.12	.06	-.001
Keep informed about politics	.27	.35	.32	.34	.29		.40	.42	.51	.23	.28	.05	.02
Engage in political discussion WHITES	.15	.22	.30	.45	.34	.52		.64	.58	.39	.34	.12	.04
Discuss politics between elections	.17	.22	.24	.38	.30	.48	.64		.46	.43	.32	.06	-.05
Inform others about politics	.15	.23	.36	.51	.46	.42	.54	.41		.32	.27	.16	.07
Tried to influence political decisions other than by voting	.16	-.27	.21	.29	.21	.22	.32	.30	.32		.35	.13	-.03
Have been asked for advice about politics	.14	-.06	.22	.35	.28	.25	.34	.33	.36	.38		.04	.01
Join in public street demonstrations	.01	-.005	-.02	.14	.08	.12	.19	.14	.14	.15	.13		.46
Riot if necessary to get political wrongs corrected	-.14	-.10	-.06	.009	.04	-.02	.03	.02	-.02	.00	.03	.41	

*Correlations above the diagonal are for Blacks; correlations below the diagonal are for Whites.

One achieves a better understanding of the political behavior of Black Americans by studying both their conventional and unconventional participation in juxtaposition to that of Whites. Despite a greater propensity to engage in unconventional behavior, Blacks in our community participate equally as much in conventional ways as do Whites. In terms of adopting socially acceptable and approved methods of expressing political preferences, then, Blacks are equally as supportive of the system as Whites. To the extent that the recent emphasis on urban violence does not touch on this point, we may be getting a one-sided view of the current effects and possible consequences of Black political activism.

FOOTNOTES

* This paper draws on experience and data derived from an ongoing study of agencies of social change and political behavior in Buffalo. The study has been supported by the State University of New York and the Office of Economic Opportunity in Washington, Contract #50-6056-A. Computing time and facilities were contributed by the Computing Center of the State University of New York at Buffalo which is partially supported by NIH Grant FR-00126 and NSF Grant GP-7318. We gratefully acknowledge the assistance of James Hottois and Ronald Johnson in the preparation of this paper.

1. Gunnar Myrdal, An American Dilemma (New York: McGraw-Hill Paperback edition, 1964) 2 Vols.

2. Report of the National Advisory Commission on Civil Disorders (New York: Bantam Books, 1968).

3. Peter Bachrach, The Theory of Democratic Elitism: A Critique (Boston: Little Brown and Company, 1967).

4. Ibid. pp. 8, 32.

5. Bernard Berelson, "Democratic Theory and Public Opinion," Public Opinion Quarterly, XVI (Fall, 1952), pp. 313-330. Reprinted in B. Berelson and M. Janowitz, Public Opinion and Communication (New York: The Free Press, 1966) 2d. Ed., pp. 489-504, p. 492.

6. Ibid.

7. Mill's argument is presented in Theory of Democratic Elitism, p. 4.

8. Ibid.

9. Ibid., p. 101.

10. For the distinction between manifest and latent functions see Robert K. Merton, Social Theory and Social Structure (Glencoe: The Free Press, 1957) Revised Ed.

11. Donald R. Matthews and James W. Prothro, Negroes and the New Southern Politics (New York: Harcourt, Brace and World, 1966) p. 37.

12. Ibid., pp. 52-58.

13. Lester W. Milbrath, Political Participation (Chicago: Rand McNally and Company, 1965) p. 18.

14. L. W. Milbrath, E. F. Cataldo, R. M. Johnson, L. A. Kellstedt, "Card Sorting as a Psychometrically Clean Device in Survey Research." Paper presented at the meetings of the Southern Political Science Association, New Orleans, La., Nov. 1967.

15. The following tables report the percentage distributions for Blacks and Whites in our sample on education and income.

EDUCATIONAL LEVELS, BY RACE

Educational Level	Whites	Blacks
Grade School	22%	25%
High School Dropout	24	36
High School Graduate	26	24
College	27	14

Footnotes (con't)

INCOME LEVELS, BY RACE

Income Level	Whites	Blacks
0 - $ 3,000	17%	25%
3 - 5,000	16	30
5 - 7,000	29	30
7 - 10,000	21	12
over 10,000	17	3

---

16. Buffalo was one of the eight cities experiencing a "major" disorder in 1967 according to the U.S. Riot Commission. Report, pp. 113, 158.

17. Supplemental Studies for the National Advisory Commission on Civil Disorders (Washington, D. C.: U.S. Government Printing Office, 1968).

18. Milbrath, op.cit., pp. 53-54.

19. For Whites the correlation between income and education is .48; for Blacks it is .36.

20. Milbrath, op.cit., pp. 56, 79.

21. Close to half of the Blacks in our sample were highly cynical, while only about one-third of the Whites were.

22. Matthews and Prothro, op.cit., pp. 78-82.

23. Hadley Cantril, The Pattern of Human Concerns (New Brunswick, N.J.: Rutgers University Press, 1966).

V

ELECTION-NIGHT FORECASTING--PANEL DISCUSSION

Chairman, LOUIS H. BEAN, Consultant, Washington, D. C.

ELECTION FORECASTING BY ABC NEWS

Richard C. Taeuber
Leasco Systems & Research Corporation

In their December, 1964, critique of that Fall's election night programs on network television, TELEVISION Magazine commented that "The computers, standing shoulder to shoulder at the television networks' nerve centers on election eve were so packed with information it hardly seemed necessary for anyone actually to vote. The machines probably could have determined the whole course of the Presidential election from the raw data supplied by one man's ballot, preferably a high-income Negro Catholic living in a split-level house with two and three-tenths children in a Midwest Polish neighborhood that went for Alf Landon in 1936."

Since the first useage of computers in 1952, the election eve coverage has been possibly the greatest single public exposure of working computers. Their use in assisting in election night forecasting and analysis is an application unlike no other in the computer business, in that this application calls for a fully reliable operation, and no one or two week postponement of election day is possible if a program bug occurs. With the heavy involvement of computers in election coverage, it must be emphasized, indeed it cannot be stressed too strongly, that computers do not vote, they do not determine the outcome of any race, they do not merely total votes, nor do they guess at the projected results. Computers do assist the human analytical staff in comparing present happenings with past occurrences, and projecting the final outcome of this particular election. Technically speaking, if the coverage was of but a single race, there would be very little need for computers. However, with the network analytical staffs trying to keep track of what is happening in the Presidential contest in 50 states and the District of Columbia, in senatorial races in 34 states, gubernatorial races in 21 states, and in 435 congressional districts, the need for electronic assistance is evident.

In the projection of election results, the analytic staffs, served by computers which use procedures supplied from the minds of mortal man, face an interesting accuracy-tolerance problem in the calling of the winners of the individual races. In a race where one candidate receives 65% of the vote, then accuracy within 10% is more than sufficient to allow a correct call of the race. If that winning percentage slips to, say 50.5%, in a two-party situation, then accuracy within six-tenths of one percent may not be sufficient to prevent calling the wrong winner, although, from a statistical point of view, such accuracy (0.6%) would be really quite remarkable for the amount of information generally available. Although the emphasis in election eve forecasting in earlier years was on projecting winners, the coverage on all three networks now has progressed to the point where in 1968 not only are the winners forecast, but also the vote-split, the plurality or margin of victory, and the total turnout. Further, the various why's and wherefore's of a particular victory will be analyzed and discussed on the air to a greater extent than ever before.

Projection Model

With the threefold replication of this single application by the networks, one might raise the question of why the projections and analyses are not handled on a pooled basis by the three networks. Although in some sense the techniques employed by the three networks have similar ingredients, the merging together of these basic ingredients is handled quite differently by the three television networks. How, then, do we at ABC make the projections? We use a projection model which is based on a weighted combination of a baseline (or time zero) estimate, key precinct information, and raw vote information. Mathematically, the general form of this model can be expressed as

$$P = \alpha B + \beta K + \gamma R$$

where $\alpha + \beta + \gamma = 1$
P = overall projection
B = baseline estimate
K = key precinct estimate
R = raw vote

The model is deliberately kept simple because we feel it important that the analytical personnel be able to understand not just the final numbers that come out but also the individual components of the model. The actual model is based on what we refer to as deviation analysis or swings, meaning that the model works with the difference between historical data and present day happenings, rather than with the absolute levels of vote being recorded in those reporting units incorporated into the model. The use of swings rather than absolute levels of vote may at times pose a problem for the analytical staff in years such as this, for there are no recent presidential races with three candidates and thus no three-party historic data is available at the precinct level. To cover a three or four-party race, then, a much greater burden is placed on the humans who are interpreting the computer output and thus, again, the need for a comprehendable model. It might be added that from the very beginning of the planning for the 1968 coverage, all systems planning has provided for handling as many as four candidates in any given race, and the move above two candidates really presents very little problem for the computer aspects of this forecasting exercise.

In the model described earlier, the weighting coefficients -- α, β, and γ -- are computed to reward consistency in that they are inversely related to the variance of the information element with which they are associated.

They vary over time as the various forms of reports reach the ABC studios, but consistently total to one. At the beginning of the evening, the total weight is on the baseline for neither key precinct information nor raw vote information is available to us. As the evening wears on, the key precinct information is the first to come in and that weight begins to climb, by design never completely supplanting the baseline. Later in the evening, as the raw vote begins to come in, that weight climbs and the weights on the baseline and key precinct information begin to decline. Finally, in the model that we use, at the end of the "evening", the coefficient for the raw vote information goes to one. Thus, at the end of the evening, if the raw vote tabulation has also been completed, our model is in agreement with that raw vote tabulation. It might be added, in a parenthetical note, that such a system presented problems in 1960, for in that year in California the "final" tabulated vote showed Kennedy with a slight lead over Nixon in the Presidential contest. However, California, in that year, tabulated absentee ballots some two weeks after the election. The absentee ballots were so heavily for Nixon as to swing the victory in that state's Presidential contest from Kennedy to Nixon. It must be noted that California, in 1968, will tabulate its absentee ballots on election night, as will most other states.

Let us look into the three aspects of this model individually.

Baseline Estimates

The baselines are really a time zero projection based on any and all prior information available to the ABC staff. They are, in a sense, the best subjective forecast by the projection staff based on polls, on educated opinions, on informed judgments, and any other sources of information available to the network staff prior to the receipt of actual vote information on election day. The estimates are numeric in nature and give the exact projected vote split, not merely an estimate as to who will be the winner.

Key Precinct Estimates

Key precincts are the primary information source on which all three networks base both their projections and their analyses. The term "key precinct" refers to precincts which are selected by one of a variety of means, and then staffed by a network representative who phones the results directly to the computer center. This provides information that can be received and analyzed as a sample of the total vote prior to its inclusion in the normal collection process. The key precincts are selected on a probabilistic basis, either completely randomly, or they may be constrained in their selection to represent various strata or ethnic groups. For the 1968 coverage, ABC selected its precincts in a two-stage sampling process. In the first stage, communities or sub-elements of the state were selected with probability proportional to the voting age population of that sub-element of the state. Within the communities, precincts were selected randomly but were oversampled. The oversample was then weeded down based on considerations of historic performance, on the availability of historic data, and on the potential availability of the results election night. If the vote from an individual precinct would not be available within a reasonable time frame on election night, if at all, that precinct was discarded from the list of key precincts. It does a forecasting model little good to receive the report of the key precinct after 80 to 90% of the raw votes of the state have been tabulated. The precincts, after being selected, are researched to give both their historic performance in several immediate prior elections and also their ethnic composition. This latter information will be used in the portions of the computer program which project and help analyze the whys of the election and how the various ethnic blocks are voting or not voting. The need to obtain historic information also provides certain problems with precincts whose boundaries or whose composition, in any one of a number of senses, have changed since the preceding election. For example, the precinct may have new boundaries, it may have a new high rise apartment, or it may have been partially cleared for redevelopment within the last two years. Because of the mobility and dynamic changes which are so widespread in this country, one cannot simply rule out those precincts which have undergone such changes, for to do so would eliminate a very major portion of the electorate from objective consideration in the projections.

But the statistical selection of the sample precincts is only part of the battle. These precincts must be staffed to report quickly and directly to the studio. Here there are problems! ABC uses a staff of one or two on-site persons to handle each key precinct. This staff is generally supplied by the League of Women Voters in each state. They visit the precinct several weeks in advance to determine, among other things, the availability of a phone for quick reporting to the ABC studios. If there is no phone available for use in the immediate vicinity, then ABC has a private telephone installed for use of the reporting team. In one instance, in 1966, even this presented a problem, for the precinct was in a rural farm house and the farmer would not permit a phone to be installed in his house. The resourceful team members looked around outside, but they could not find a suitable building in the vicinity. They did notice a very large oak tree, and, following a request to the telephone company, a phone was installed in the oak tree. There are also additional problems that the field reporters face, such as a family of skunks under the floor of a precinct polling place; precincts in which only Spanish was spoken; fuses which blew; fire in telephone company offices; and, in more than one instance, teenagers using the public booth telephone who may still have been on the line the next morning. The reporters also performed a variety of chores, including providing dinner or snacks to permit election officials to get on with the ballot counting without taking a dinner break. And the reports from 1966 also include comments on the problem of how to prevent absentee ballot-box stuffing and observed instances of unlocked ballot boxes.

Raw Vote

The final component of the model is the raw vote as it is reported by the News Election Service. The News Election Service is a stepchild of the three networks and the two main wire services. It was started in 1964 when it became very obvious that a five-fold tabulation of the raw vote made little practical sense and could be done much faster, more completely, and much more economically on a pooled basis. This pool in 1968 will receive reports from about 80% of the precincts in the United States as well as from the county seats of all the counties in the United States. The tabulations will be performed by computer for all 50 states, and the source which provides the most vote, at that moment, that is, either the NES tabulation of precinct returns or the county reports, will be forwarded to the members for broadcast or relay to the viewing, listening, or reading public.

The incorporation of the raw vote into the projection model calls for an awareness of the patterns of reporting within many of the states. For instance, in New York State, it is well known that New York City is reported before the rest of the state, and that it casts a higher proportion of Democratic votes than the upstate portions of the state. There is a similar well known pattern in Illinois, with Chicago or Cook County reporting earlier and being more preponderantly Democratic than downstate Illinois. Similar patterns, moreover, hold in varying degrees in a good many states where the early vote may be more Republican or more Democratic and the discrepancy may disappear monotonically, or may, in fact, swing the other way during the middle of the tabulation process, and then later disappear. It might be noted parenthetically that the faster tabulations which will be produced in '68, with computer assistance, add an additional element of uncertainty to the accuracy of these historic reporting patterns. One way to minimize the effect of this uncertainty in reporting patterns would be to use finer geographic breakdown than the state level to permit the model to compensate for any such pattern. Such segmenting is being done, but the timeliness of the availability of aggregated data may force whole-state useage at various times throughout the evening. When this occurs, a correction factor may be entered to compensate for the historic patterns. Such a correction factor would normally be expected to go to zero at the end of the evening. However, one might refer back to the earlier comments about California in 1960 to note that perhaps this correction factor should not go to zero in certain states. Although the computerization of the election tabulations and the movement throughout the country towards nonpaper ballots (and thus faster counting) may disrupt the patterns of statewide reporting, the computerization does provide timelier availability of the data which permits the use of finer geographic breakdowns, even to the county level.

In closing, one might draw an analogy with weather forecasting. Both weather forecasting and election forecasting are done by technically competent staffs which have to make allowance for many factors outside of the control of the projectionist. One case leads to ruined picnics; the other case may salvage a supposedly "ruined" evening by providing the hope that the forecasters on television have "goofed" in some manner. The weather forecasters receive a bit of tolerance for their prognostic inaccuracies and are judged by their peers for their interpretation of the available evidence. Unfortunately, those who make judgments as to the competence of the election forecasters do not have knowledge of the totality of the available evidence, or even an indication as to what that available evidence is. Election projectionists are allowed little leeway for error in their interpretations. The volatility of the American voter continues to present a problem to those who are involved in election forecasting, but the credibility of the projections has run high in the past and is increasing.

CBS NEWS -CAMPAIGN 68

WARREN J. MITOFKSY - CBS NEWS

CBS News will report the results of the 1968 elections for each state-wide presidential contest and all gubernatorial and senatorial contests. We shall also analyze the source of the candidates' votes and compare them with past elections at state, regional, and national levels. The source of a candidate's vote will be estimated by geographic area; by size of city and rural areas; for areas classified by income, occupation, past voting behaviour, and various other characteristics by which geographic areas can be classified. We shall also attempt some measurement of the issues influencing people to vote as they do. We shall try to separate, in our analysis, the salience of issues as they influence change in voting behaviour from the past. It may be somewhat obvious, but let me point out the source of the various analyses. The estimated vote for various classifications of geographic areas will be based on samples of precincts. Actual returns on election night will provide the source of these data. Any estimate related to issues will, of necessity, be the result of pre-election sample surveys. In addition to state-wide elections, we will estimate, by geographic region, the party composition of the House of Representatives. If the election for president is not to be decided in the Electoral College, where an absolute majority of electoral votes is necessary, the election will be decided by the House of Representatives, where each state has one vote. If this contingency seems likely, we shall report the majority party for each state in the House of Representatives.

Our computer hardware, supplied by IBM, consists of two 360's, model 65, each with one million bytes of core. All programs, past data and election night data will be retained in core and it will not be necessary to access disk or tape storage for any calculations. On election night, all input will be by use of IBM 2260's, which will eliminate any use of punch card equipment. For those of you unfamiliar with this device,the 2260 is a display screen with the ability to enter data or recall data from a computer. Another device, the IBM 2250, will be used for graphic display and display of all estimates of candidates' percentages on air. The programming is being done by Informatics, a company which has extensive real-time programming experience on such projects as the airline reservation system and the Western Union message switching system.

In 1966, for the first time at CBS, election estimates were based on a probability design. This year we hope we have made some significant improvements both in procedures and use of available data. This includes use of both NES and returns from samples of precincts.

The estimation of the outcome of an election is treated in many ways as a traditional sample design problem. A single stage sample of precincts is selected with probability proportionate to size from a stratified frame which includes data from a past election for all precincts in a state. The stratification is based on past vote for precincts as well as county characteristics. Various estimators are available that will make maximum use of the usually high correlations of total vote or party vote with past elections, such that variance can be minimized. In fact, precise estimation of election results would not be conceptually difficult if it were not for the requirement imposed by the networks for calling the winner at the earliest possible time.

Before enumerating these difficulties, let me clarify a few points that are usually raised about early calls. First, the results of an election within a state are not broadcast without some minimum amount of actual election returns being available from the state. This means that either some or all of the polls are closed within a state before estimates are possible. Second, if all polls within a state are not closed at the time a winner is announced for that state, the local stations in the state are cued and have the option of not broadcasting the result.

Third, I know of no study that has adequately estimated the number of potential voters who have not voted at the time results are announced for the east. And, of this group of potential voters, those who have been aware of the announced results, and who also have been influenced either not to vote or to vote for a particular candidate. The last part of such a study should estimate the impact of an election.

The first special statistical problems related to the early call of winners has to do with missing reports for sample precincts. The usual assumption for most statistical models is that accurate measurement exists for all elementary units in the sample. If measurement is missing for elementary units, then this component of error must be included in the determination of the mean square error. If elections are to be called as soon as possible after the polls close, with a predetermined risk of calling the wrong winner, it is necessary to measure the error due to missing returns from precincts, as more than half of the precincts may be missing at the time of a call.

If the correlations between a past and present election are high, it is possible to impute results to missing precincts. If correlations are low, it may be reasonable to assume that those precincts reporting are approximately a random subset of the initial sample. Of course, the subset would have to be relatively large. It is also possible, for defined geographic areas within a state where the reporting of precinct results is known to be slow, to make election day estimates of actual voters. These estimates may be superior to imputation based only on actual returns. In any case, it is necessary to select the better imputation procedure and estimate the resulting contribution to the mean square error.

The second major statistical problem relates to quality control. On election night it is almost impossible to determine whether results are reported correctly for a particular precinct. I might clarify what I mean by incorrect results. This happens when unknown to us, a precinct's boundaries have been changed or when precinct names have been changed. In either case, we can receive correct results for a geographic area different than the one selected in the sample. This can yield misleading results as the various estimators usually depend on the correlation of total vote, and in some cases of party vote between present election and a past election. Some control of input data can be accomplished by having various criteria for acceptance available to the input operator on recall from the computer through the 2260. Alternatively, the report may be correct but the results may act like an outlier. The recognition of "outliers" versus trends in data is a significant problem. Consequently, the acceptance or rejection of data must be flexible enough so as not to distort the estimate at any given point in time. The rejection of data should be done so that the additional bias will be less than the reduction in sampling error. An error in quality control procedure occurred for the Maryland gubernatorial election of 1966 where CBS called George Mahoney a "probable winner". The call resulted from a failure to distinguish between a trend and an outlier. Consequently, data was completely and incorrectly rejected when it should have been accepted.

The third problem has to do with the constantly changing precinct boundaries. If sample precincts are selected from either 1964 or 1966 sampling frames, it is quite likely that a substantial number of boundaries have been significantly modified. It is possible through a great deal of tedium to resolve boundary changes for individual precincts within slightly larger geographic areas that remain common over the years. An unbiased measure of size for either 1964 or 1966 total vote can be obtained for a precinct as it is geographically defined in 1968. Reasonable approximations to the party vote also can be made. Failure to recognize boundary changes during the time of the field work, results in much of the incorrect data described earlier.

Last, but not least of the problems, is the necessity of establishing criteria to exercise the proper options of quality control, imputation, weighting, estimation, and estimation of the mean square error. For example, while a regression estimate theoretically might yield the smallest variance, the problem of estimating the regression coefficient reliably might make a ratio estimate more desirable. Of course, "if the Wallace third party is at all effective ..., analyses based largely on past experience might go wrong and 'all bets would be off'," as our chairman prophetically said in his book about elections 20 years ago. In the case of base correlation of party vote with a past election, estimate will be based on current data only. Also, imputation must proceed differently when either the correlation with party vote is small, or when a third party candidate has a reasonable share of the vote.

Other criteria, related to the decision as to the winner are under the final review of a statistical team. We believe it is possible to program most of the decision process, but not all. People are still most able to recognize patterns or irregularities. The recognition of the winner of a race is the focus of the decision. The actual percentages for candidates is a by-product not subject to the same rigor as the winner decision. Elections are the only sample survey problem where three groups are doing the identical survey and parameters are available (in most states) shortly after the estimates. If the proper options are exercised incorrectly, the failure will soon be known coast-to-coast.

I might conclude by pointing out that almost any statistical design is adequate for a landslide election. Furthermore, no design will be reliable for an extremely close contest. It is the middle ground, where elections are won by 2 to 10 percent margins, that the networks' election estimates have their greatest efficacy for reporting results in a timely and informative way.

VI

IMPROVING FEDERAL STATISTICS ON CRIME AND CRIMINALS

Chairman, DANA M. BARBOUR, U. S. Office of Statistical Standards

IMPROVING CRIMINAL STATISTICS

Leslie T. Wilkins, University of California

The title given for this morning's symposium seems to imply that we are of the opinion that criminal statistics can be "improved." I hate to start so early in the morning with a semantic point; but clearly it all depends upon what we mean by "improved."

Let me make my own position clear, even at the risk of offending. It is my considered opinion that problems of statistical data regarding crime, criminals, dispositions and outcomes of offenders, not to mention victims, are deep-seated. I do not think that minor adjustments of the classifications and new forms of processing alone can provide what is required. The problems are fundamental to the whole conceptual framework of the provision of statistical data which relate to the purposes of the administration of justice and the treatment of offenders. There is no stage in the network -- from the first initial contact between the suspect and the law officers, through to the discharge on parole and surveillance of the person "with a record" -- which should not be given very close attention and thought through in the light of modern social science.

COMPUTER TECHNOLOGY

The availability of high speed data processing equipment changes only the less difficult aspects of the problem of the provision of adequate data. There are many things which the present generation of computers cannot do. Too often these limitations are overlooked and the computer has come to be seen by some as a new general magic -- a sort of philosopher's stone which can turn large quantities of data into solutions which will make completely clear the decisions that should be made.

The computer cannot:

(1) set the boundary conditions of a problem;
(2) define a problem;
(3) say how worthwhile it is to explore a particular problem in the first instance and hence cannot indicate the rational use of resources of money or persons in work on an original problem;
(4) imagine the variables or other information which might be relevant to a consideration of a problem;
(5) decide what to include or exclude in the initial operations on a problem (but see 1 and 2 above);
(6) select the functions to be explored;
(7) make decisions regarding the range to be covered by an included variable;
(8) construct a model;
(9) select the criteria to be explored (although we may program a computer to select the "most predictable" criterion of a number which we previously have noted for inclusion);
(10) decide how to collect the basic data;
(11) decide upon a strategy of operations, such as deciding what proportion of resources should be devoted to different stages of a problem, or the means for its solution;
(12) design or evaluate a sensitivity analysis.

If we are to treat the problem of the form which criminal statistics should take as a research question, it is necessary to transform the question into operational terms. One of the first steps is to set down some boundary conditions.

BOUNDARY CONDITIONS

Presumably we are interested in "crime" for a variety of reasons, but mainly because we are interested in information as it is related to social administration (i.e., control) systems. Such types of information are different from those which might satisfy "idle curiosity." Of course the scientist should be curious, but scientific curiosity is different from that of those concerned with problems of social policy. Let us begin by limiting our consideration to criminal statistical data in relation to questions of social policy. We will not be satisfied if, in reply to our question, "Why do you want to know that?", we can get only the reply, "Wouldn't it be interesting?" In other words, there must be some referent to a purpose other than the personal interest of the individual. I would suggest that this basis of reference might best be stated in terms of decision alternatives. If there is no possibility of any variety in the decisions which must be made, then there would seem to be no purpose in seeking data. If there are decision alternatives, then we may consider what would be the nature of the information which could inform regarding the selection of decisions among them. Moreover, if the same decision would be reached given an estimate of $x \pm 1\%$ and an estimate of $x \pm 10\%$, then the cost (and there must be such a cost) of reducing the error from 10% to 1% is a waste.

THE NATURE OF THE PROBLEM

In the time available to me, I can select only some of the major areas and note some of the problems as I see them. I shall propose no solutions, but I hope to be able to indicate some ways in which the problems I shall raise may be approached for investigation.

It is my view that in considering the matter of large-scale information and organization, such as crime statistics, we cannot expect results from methods of inquiry which differ significantly from those which we have used successfully in small-scale problem solving. I refer, of course, to those means which have been termed "the scientific method." I do not believe that we can sit down and think ourselves through to solutions as did the old time philosophers. Informed opinion is not adequate; there must be sought the various forms of evidence which will enlighten the problems we can describe. Our first task is to specify the problems. The next is to consider what kinds of data will assist us in their examination. Changes which may be made in criminal statistical data which are not themselves based upon sound research rather than opinion (even expert opinion) are not likely to be sound. I would also claim that it is probable that once we approach the revision of criminal statistics from this scientific viewpoint we shall find that much activity which is necessary will fall within that class of research methodology that is loosely termed "fundamental research" -- applied and even operational research will not be adequate. But let me try to establish the case by some examples of problems. Let me start with some easy examples -- easy in the sense that they can be fairly clearly specified.

Age and Sex Factors

Everybody knows that the majority of reported and cleared-up crimes are committed by young male persons. Similarly, everybody knows that the probability of death increases with age. It has long been recognized that death rates for cities or other areas or classes of persons are not open to any useful interpretation unless they are age-standardized. Given the crime data which we have available, it is often possible to make some analyses of an age-standardized type; but, of course, such data can relate only to those crimes which are related to a person. Strictly speaking, such rates are not "crime" rates; they are a kind of "person-decision-event" rate.

For all we know about "crime" at the present time, there is bound to be a time-lag correlation between the birth rate and the "crime" rate. This seems all too simple, yet I have seldom seen this factor taken into account when "crime" data are published or even subjected to analysis. But clearly this factor is only one of a type which we may relate to measures of "exposure to risk."

There are certain forms of inference which may legitimately be made on the basis of the crude data, but there are many other forms of inference which we need to make that require a much more sophisticated index. Any inferences regarding the "state of crime" which may be made on the basis of "persons-in-respect-of-whom-a-decision-was-made-to-make-an-arrest" are unproven.

Exposure to Risk

In analysis of road accident data, much attention has been paid to the problem of obtaining a reasonable estimate of factors which reflect to some degree the various aspects of "exposure to risk." It has been realized in studies concerned with traffic that few sound decisions of policy can be made without some such base to which to relate the crude figures. In this case it is natural to think of such data as estimates of mileage. But what are the factors which it might be reasonable to consider in relation to the "crime" data? And, in any case, what exactly do we mean by the "crime" data? If we consider murders, or with somewhat less justification offenses falling into the category of "crimes against the person," it might be reasonable to use the population as a base for "exposure to risk." But is it equally as reasonable to use the population of persons as a base for offenses against property? If the cost of living (legally) goes up, presumably the cost of illegal living also rises! If persons who live by illegal means increase their productivity proportional to the increase in the cost of living, does this really mean that "crime" has increased?

I have shown elsewhere [1] that for certain offenses where the opportunity (or exposure to risk) increased, the offenses committed followed a very similar pattern over time; indeed the ratio established over a period of twenty-four years (1938-1961) remained almost constant, despite fluctuations in the absolute figures of 800 percent. When there was less in the shops, less was stolen; when there were fewer cars on the road, fewer were broken into. This analysis was necessarily crude because economic data and "crime" data cannot be matched. Larceny from shops and stalls, which was one of the categories used, would exclude "burglary" and probably many or most cases of "breaking and entering" of shops, but it seems difficult to me to commit burglary or breaking and entering in respect of a "stall." Perhaps this example serves best merely to illustrate a point, namely, that illegal economic behavior (crimes against property) cannot be matched and compared with legal economic behavior of an otherwise similar type. I see no utility in data which cannot be compared with other data -- I am not that kind of believer.

In the case of murder, it seems fairly clear that "persons" are at risk either to becoming murderers or victims, but there would be a different rate if we used victims as the base, since murder is not a one-to-one transaction. In the case of property offenses, the connection between the person (owner of the property) and the exposure to risk is distinctly less direct; much property is owned by collectives of persons. Perhaps fluctuations in appropriate insurance rates for certain classes of risk would provide a better indicator of some categories of "crimes" than the data which are available from "criminal statistics." At least, with such data there is a basis of exposure to risk which, in the experience of the companies underwriting the policies, provides guidance for the setting of premium ratings, and we may assume that competition between companies will ensure a good market.

The Concept of Risk and Probability

It must be nearly one hundred years since it was considered to be good statistical practice merely to count things -- a sort of numerical accountancy. Statistics are concerned with probability and decision and related matters. If, then, we use the term "criminal statistics" we might expect to find, at least as raw material, data which could be stated in terms of probability estimates. But can we? Hardly at all. We might think that if the word "criminal" has any meaning we should be able to know how many "criminals" there are. But we cannot obtain even a good estimate of the many possible definitions of that word. If we were to take a random sample of the population, what would be the probability of sampling a "criminal?" Is even that concept sound?

Perhaps even more important, how many "victims" are there in different classes and under various possible definitions? Here we are in a worse state of ignorance. We cannot inform a citizen of the general expectation of his suffering any particular attack upon his property or person. Is there such a factor as "victim-proneness?" Would it not be useful to know? Clearly, without a measure of "exposure to risk" related to the "crime" data in each case, we cannot utilize statistical methods at all -- that is to say, we cannot make reasonable estimates of the probabilities which form our stock-in-trade. The foregoing explains my reason for placing "criminal statistics" in quotes; without an estimation of probabilities we do not have statistics. But why also place "crime" in quotation marks? We may think that there is no doubt about the fact that we do have "crime." I would agree that, as a statement of a social condition, this statement is a good one. But our data do not relate to crime.

What Data Do We Have Which We Call "Crime Data?"

When is a crime not a crime? If we are uncertain about even this dichotomy, can we be very precise about measurement? Clearly, the only crimes about which we can have data are "known crimes." But we have no means of knowing by direct revelation; we have to carry out processes and note the results of the description and classification of these processes. A cursory examination of "criminal statistics" will quickly reveal that we have certain information about:

(a) persons;
(b) events;
(c) decisions;

but in general we have a compounding of two or more of these. In the main we have data relating to decisions -- the policeman decided to arrest, the judge decided to find the person guilty and to make a certain disposal, the parole board decided to parole or to recall from parole, and so on. These decisions have a relationship to an event which we have classified as a "crime," but the relationship is by no means a one-to-one relationship. The "event" which is classified as a "crime" is, in almost every case, not a single event, but is better considered as a transaction. The criminal cannot generally commit the crime upon himself.

I tried recently [2] to make a short statement which reflected the complexity of the problem of criminal behavior. It may not be adequate, but perhaps I can repeat it here as an illustration of this point:

> Persons who vary in ways which are in the main unknown (X_1) live in situations (X_2) and are exposed to cultural influences which vary in unknown ways (X_3), they sometimes commit deeds (X_4) which vary in many ways except that they are classified by the laws of that society as crimes, and these laws (X_5) also vary both in content and interpretation; some persons are detected by systems which vary in unspecified ways (X_6) and are dealt with by persons or courts which also vary in their policies (X_7) and are allocated to institutions or disposed of in other ways usually termed "treatments" (X_8), which differ from each other in unknown ways. The persons are committed for varying periods of time (X_9) and their interaction with the "treatment" (X_{10}) will be expected to vary; in most cases they may be expected to interact with other persons (X_{11}) also undergoing the "treatment." Eventually they are released to situations which vary both in themselves and in terms of the expected interaction with the personality of the ex-inmate (X_{12}).

The twelve variables (and there may be more) would be repeated for each offense where more than one is concerned; that is to say, for most persons who are identified as criminals by our present processes.

It will be noted that the set of variables outlined above considers the process of criminal justice mainly from the viewpoint of the offender and his rehabilitation. There is another viewpoint, namely that of society, and the impact of crime upon the general social condition -- we may call this the macroscopic viewpoint. There are, of course, also other viewpoints between these two extremes. If we are concerned with rational processes then there must be as many networks of information as there are networks of decisions in regard to those processes which have been constructed or have grown up over time to deal with issues of social control.

Clearly the problem is complex. Yet it is strange how many persons, often in high office, see the problem of crime as a simple matter. There is for many still a definite line between right and wrong, and a two-value logic (and not always logic) suffices as an explanation which they find satisfactory; at least it seems to give them satisfaction.

An Analogy

Consider for one moment the very large body of data which are available to economists. (I will not comment upon how well they are able to utilize it!) Now these data, by definition, concern legal economic transactions. Crimes, also by definition, are illegal and frequently economic transactions. (I maintain that we cannot use the scientific method if we consider crime from the moral viewpoint.) Now legal behavior follows laws; thus the variety of behavior is constrained and hence more predictable. The regularity which may be found in illegal transactions is a "law" of a different order, and the process may be expected to be more complex. If this argument is sound, then it would appear that for purposes of examination of illegal (unconstrained by law) behavior we may need more complex data than we need to permit examination of legal economic behavior, and illegal economic behavior is only the larger part of crime. I do not want to press this point, but it serves to indicate that no easy solutions are likely to be found.

NECESSITY OF SEQUENTIAL ANALYSIS

It is, I think, a pity that the advent of the computer has put sampling strategies somewhat in the shade. It is perhaps time to revive considerations of sampling in our strategy of information collection, processing and analysis.

It is not possible to ask all the questions which we will want, at some time, to ask about problems of crime, criminals and the processes of justice. Answers to questions provoke more satisfactory questions. We should not attempt an "ideal" solution, but rather set up a system which can be continuously modified as our ideas of appropriate questions change. I think we can get from economics some guide lines to the methods which might be useful. In economics macro and micro methods are well distinguished. So far there seems to be no similar distinction in the thinking or the data regarding illegal transactions. Nonetheless, similar distinctions will eventually have to be made, and different types of models will have to be applied to the two frameworks. The setting of rational budget allocations for crime prevention and control would seem to require models similar to those provided by macro-economics, while evaluation of penal treatment/punishment probably would require being approached by the "black box" model.

Decision "Gates"

The judicial system can be pictured as a network of decisions -- rather like a branching tree, with each branch having a "gate" (decision). It would appear to be an easy matter to fit each of these gates with a simple counting device. At present only a few of these gates are covered by counting systems, and I doubt that all the gates have been systematically identified. Where we have such "gate counting," we do not know how many of the persons passing through any one gate have passed through other types of gates previously on their route through the system.

Adequate counting systems at each of the decision gates should be related to a sampling frame and to a means whereby good probability samples could be drawn as and when required. There is no need to try to solve all the problems all the time! Sampling could suffice for many of the "macro" models as well as being the general method for testing "micro" analyses.

All too often an attempt at 100% data collection falls short of a complete enumeration. In such cases it is normally better to have a good probability sample than to have empty cells in the tabulations. In a recent study I found an almost complete enumeration (92% of the universe of interest) totaling about 20,000 cases. It was believed by the office concerned that the missing data were not biased. Nonetheless, I decided to utilize high order interactions as an empirical estimate of error. It was found that the power of the 92% enumeration was approximately equal to a random sample of only 2,000 cases. This was due to the increased error variance of non-response. In this case it might reasonably be inferred that the agency concerned was wasting 90% of its effort in collecting 92% of the data, since a 10% sample would have been equally as efficient as the attempted complete enumeration, and would certainly have been less misleading.

NEED FOR SPECIFIC DATA

Discussions of general statistical data seem to assume that if only one could collect "sound" and appropriate data, such data would serve all kinds of purposes. I do not take this view. Specific needs require specific types of data. The publication in the press of "criminal statistics" does not, I fear, inform the public. The perception of "crime" by the average layman does not coincide with the technical legal definitions. The layman may be pardoned if, when he hears about the annual increase in "crime," he places a summation sign in front of the various news headlines which he has seen and can remember. Headline or, indeed, newspaper reported crime has specific characteristics related to the selling of newspapers rather than to the control or treatment of offenders.

If we could tell members of the public what was the probability of their being the victims of various types of crimes and if we could amplify this by indicating different classes of risk situations, they would be better informed and able to adjust their behavior accordingly. The information we give the public at present does not enable them to do anything, except become more fearful in a vague sort of way. Ill-informed behavior may, by indirect means, serve only to provide a situation which makes crime more probable or profitable.

The provisioning of prisons, estimation of future inmate populations and many other administrative decisions require data each of a specific kind and adapted to the specific need. Many decisions are made at present on the basis of what Huff in his delightful little book, How to Lie with Statistics, calls the "semi-attached figure." [3]

Doubtless, much data could be obtained as a by-product of the ongoing decision processes (the "gates"). It would seem to be useful to survey the kinds of information which are at present available and used (or said to be used) as a basis for current kinds of decisions. But first we should have to map out the decisions. I do not know of the existence of any such complete decision map. Without knowing what decisions are being made, how can we discuss what information would the better inform those who make them? Moreover, it is not always those who are held responsible for the decisions who in fact make them. The higher in the hierarchical decision-making structure a person is, the smaller the proportion of "his" decisions which are in fact made by him; rather he is a decision ratifier who takes responsibility for his ratification. In a recent study with my colleague Carter [4], I have shown that judges in the Federal courts (surely persons who usually are assumed to make decisions) may more reasonably be seen as mainly ratifying decisions -- the "decision" is mostly contained in the "recommendations" of the probation officer. To what extent we consider a recommendation (which is more often than not accepted) as a "decision" or the ratification as the decision, depends upon semantics rather than a model of the decision-making process, but it is a key issue in considering who should have what kinds of information.

THE LAST QUESTION

What information should be available? This is the final question, not the initial question. What decisions can be made (what are the decision alternatives)? What information may relate to these decisions? What is the pay-off sought? What is the probability of pay-off from various decisions? These are more primary questions. I would contend that we cannot get anywhere by asking the last question first. Statistics is a method, not a belief system. Statistics can be used in studies concerning agriculture, astronomy to zoology, zymosis and many things in between, but statistics is not agriculture, astronomy or anything else. Criminologists, lawyers, sociologists, administrators and all others concerned in whatever manner with what they term "crime" cannot look to statisticians to tell them what data they require, without their first telling the statisticians exactly what are their concerns. Even when the concerns are expressed, they must be in a form related to "rational" decision processes. This is, perhaps, the major problem.

RATIONAL DECISIONS?

I suggest that many of the procedures which are followed in relation to behavior which is classified as "criminal" are probably not even intended to be "rational." Certainly not "rational" as the statistician may use the term.

Statistical methods cannot inform the artist or musician (except perhaps the composer of electronic music) and there are, I think, many situations in the operations of the processes of justice which are more analogous with an art form than with a "rational process." To say that the courts are not "rational" may be considered both incorrect and somewhat irreverent. I do not have any intent to be irreverent; rather, the perception of the judicial process as an "art" seems the more closely to accord with the ways in which some of those closely concerned in the process see their function. Fink [5], for example, specifically rejects mathematical estimates of recidivism and says that he "is inclined much more to the judgment of a judge who is wise, humane and just than to the efficiency of prediction tables." The "human touch" is greatly respected in the assessment of the work of the courts and the "human touch" is, I think, an "art form." But the "art" aspect of the work of the courts goes much further. Imagine the courts stripped of all their ceremonial and symbols; would not justice then be seen as a rather different thing? [6] Hardly a church exists without ritual and some pomp and circumstance. Even the military machine utilizes the parade. How would we rationally and statistically assess the flag? Could there be justice as we know it without drama? Is the drama not part of our very concept of justice? If so, then we must ask how rational decisions mix with drama.

Information regarding crimes and the judicial process which reaches the general public through the mass media is frankly stated to be "drama." "DRAMATIC SCENE IN COURT," runs the headline. But the dramatic event is also a rare event. That which happens every day and everywhere is not dramatic in the way in which the newspapers use the word.

The medieval morality play provided a form of social information and control for the layman and, I suppose, looking back on those times we may consider that it was a useful piece of symbolism. We have not attempted to collect statistical data with respect to prayer and sacramental performances in our various religious organizations. Statistical data can tell us how long candles of different types may be expected to burn, but we cannot assess the impact of their burning upon the worshippers or the immortals their burning is supposed to impress. In many respects, today the courts are providing a morality play for the information of the public.

As statisticians, however, we would seem to be committed to the belief that the dramatic incident is not a valid guide to public policy decisions. This is a belief which emerges from our discipline. People with other disciplines (or lack of disciplines) may have other beliefs.

The value of the burning of candles is today in doubt. The value of the operations of the courts in the role of safeguarding the necessary functions of society is also coming into doubt. The moral absolutes based on beliefs no longer underpin our social system with a firm consensus. Yet we wish to find a means for the operation of those social controls which are essential to the development of man. Our questions should encompass not only the phenomena which we have come to term "crime," but should extend to investigations of morals and public policy. The data we may obtain from the operations of the various aspects of the judicial processes may be useful raw material for the study of some aspects of the social control system. But the judicial system is only a very small part of the total social control system which has developed as man has attempted to live in societies with other men, and the learning process has been slow.

The priority in criminal statistics is to provide a basis and a set of references for the study of social control processes, and this study is, I think, best attempted at this time by sampling techniques related to basic research. Perhaps such research could reveal the questions we should be asking. At this moment I cannot claim to know even the questions.

I look forward to enlightenment from other speakers this morning.

REFERENCES

[1] Wilkins, Leslie T. Social Deviance. Prentice Hall and Tavistock, 1965.

[2] ________. Evaluation of Penal Measures. Random House, 1968.

[3] Huff, G. How to Lie with Statistics. London: Gollanz, 1955.

[4] Carter, Robert and Wilkins, L. T. "Some Factors in Sentencing Policy." Journal of Criminal Law and Criminology, 58 (1967), pp. 508-14.

[5] Fink, A. E. "Current Thinking on Parole Prediction Tables." Crime and Delinquency, 8 (1962), p. 227.

[6] Arnold, Thurman. Symbols of Government. Yale University Press, 1935.

USES OF SURVEYS FOR ESTIMATING CRIME INCIDENCE

Albert D. Biderman, Bureau of Social Science Research, Inc.

The use of surveys of random population samples for studying crime has been touted by many commentators on the work of the President's Commission on Crime and Administration of Justice as the most important recent innovation in the study of crime. I am inclined to agree with this comment, although, as I will try to make clear, I have many painful experiences with limitations, difficulties and costs of the method. That using cross-sectional population surveys should be a major innovation, however, is a commentary on the stultifying influences that have reigned in the field of criminological statistics during the past 20 years--years which have seen revolutionary developments in our capacities to develop useful and significant social information. Given what is now the overuse of sample interview surveys for thousands of trivial purposes, private and governmental, it is striking that this garden-variety means of studying almost anything in the human field was not used for the study of crime in any consequential way until 1966.

The primary effort in these studies was directed to forming estimates of the frequency with which citizens were victims of criminal offenses. The basic rationale for undertaking this work was set forth in September 1965 in a proposal made to the President's Commission on Crime in the District of Columbia:

> Criminological studies have largely developed their data from law enforcement, correctional, and judicial agencies, and from persons arrested or jailed. The only consequential exception to this are studies using high school and college students as subjects or respondents. As a consequence, there is a vast *terra incognita* in our knowledge of crime. Consider, for example, the accepted proposition that "offenses known to the police" are the "best" measure of crime because these are the data "closest to the commission of the crime." Insofar as offenses of victimization are concerned, it would seem that data developed directly from questioning the public would be "closer" to the crime and, for at least many classes of offenses, would suffer less from errors of underenumeration than data derived from reports to the police and crimes known directly to the police.
>
> Some theorists go so far as to assert that "crime" should be defined to refer to relationships between those committing certain acts and agencies of justice if the concept is to accord realistically with the data used in its study....
>
> One method of reconnoitering some of the base of the iceberg has yet to be employed in the U.S. This would involve questioning a large statistical sample of the population about direct experience with crime during a given time period. While this method would be subject to various types of inadequacies and errors of reporting--notably, insofar as crimes of self- and mutual-victimization and those in which the victim is an impersonal entity are concerned--it would nonetheless represent an enormous supplementation of existing knowledge of the extent of criminal behavior. However consequential the omissions and distortions of the images of the relevant phenomena that might be involved in the application of this method, they would in any event be different ones than those of the traditionally applied methods. As in many other problems of scientific observation, the use of approaches and apparatuses with different error properties has been a means of approaching truer approximations of phenomena that are difficult to measure.

Quite apart from the many published technical criticisms of currently used crime statistics, it was felt to be particularly essential, in view of the innovational law enforcement programs being contemplated, to develop some information regarding the incidence of crime that was independent of reports of citizens to the police. This is the case because some of the improvements in law enforcement may have the incidental consequence of increasing the number of crimes known to the police and hence estimates of the crime rate for an area, while actually these measures may reduce the occurrence of crime. A possible effect of this type can readily be discerned in the area of police-community relations. Where many of the citizens are hostile toward or fearful of the police, presumably they are reluctant to deal with them except when necessary. Under such circumstances, many crimes would go unreported. As police come to enjoy greater confidence and respect, a citizen is more likely to see some possible benefit, and certainly no harm, in reporting to the police when he is victimized.

More direct effects of improved police practices on crime statistics have been noted in a great many cities including New York and Chicago. In these cities, instituting a more professional attitude toward crime statistics resulted in extreme elevations of reported crime rates.

Heightened salience of the crime problem attendant to well publicized reform measures can also produce spurious elevations of the crime rate. For example the individual citizen's greater concern with the possibility of being victimized may lead to more widespread purchases of theft and burglary insurance. The need to make a report to the police in order to establish a claim for loss provides the covered citizen with a reason for reporting an incident to the police where he otherwise might regard doing so as completely pointless.

The interviewing studies undertaken through grants and contracts of the Office of Law Enforcement Assistance, Department of Justice, to support the work of the President's Crime Commission, included developmental and pretesting work using Washington, D.C. population samples, a survey of a national household sample by the National Opinion Research Center, and coordinated intensive studies in selected police districts in three major cities by the Bureau of Social Science Research and the Survey Research Center of the University of Michigan.

The objectives of the surveys went considerably beyond exploring the incidence of criminal victimization. For each crime incident mentioned by respondents, they were questioned on when, where, how and why the offense had taken place, on characteristics of the offenders; on the extent and nature of losses and recovery or indemnification for these; and on investigatory and adjudicatory processes that ensued. All of this information regarding victimization could be related to detailed information that was also collected in the surveys on social background characteristics of the victims, their attitudes toward many issues relating to crime, law enforcement and justice; and some other relevant aspects of their experience and behavior.

Contrasts of Survey and Agency Statistics

I cannot discuss adequately here the methods or results of these several surveys. Reports prepared for the Crime Commission are available (Biderman, et.al., 1967; Ennis, 1967; Reiss, 1967) and Reiss and I (Biderman and Reiss, 1967, Biderman, 1967) have recently published some critical examination of theoretical and methodological problems inherent in these studies. I will rather use this occasion to consider the implications of some instructive contrasts between the interviewing survey approach and the agency statistics on which our knowledge of crime incidence hitherto has been exclusively based. The survey method differs from agency statistics in the following ways:

1. Attention to events that elude agency attention and action. The primary purpose of these studies was to try to measure unreported crime--crimes that the police never learned about. The country is as interested in that part of a social problem that altogether eludes the apparatus set up to deal with it. Indeed, in some ways, there may be even more concern with what is not being dealt with at all by our agencies for controlling problems.

2. Special adaptation to information purposes. The entire system for collecting and processing information was organized and structured for meeting informational requirements and was independent of administrative or operational concerns including personnel, concepts, instruments, controls and incentives. Procedures were selected for the purposes of understanding and explaining events, rather than action, administration or case judgment. Operational definitions can be readily and truly uniform, with measurable reliability. (Actually for the work done for the Crime Commission, some definitional hobbles were placed on the procedures to make it easier to compare results with official statistics.)

3. Sampling. Probability samples rather than attempts at universal enumeration were employed. (Presumably, there is now widespread understanding of why it is that we are usually able to approach more exhaustive representation of phenomena by the sampling method rather than by attempts at total enumeration, so I need not elaborate on this point.)

4. Motivations of information sources. Where agency data collection exploits the interests of parties in the outcome of events to secure data from them, survey methods usually seek to exploit the advantages that stem from the fact that no material advantages ensue from testimony. This includes the guarantee of anonymity to respondents, the absence of sanctions or rewards to control the testimony given, and the release of aggregated information only.

5. Freedom from jurisdictional constraints. Attention can be given to whatever events and whatever aspects of them that one wishes to select, either because of their social importance or their potential usefulness for causal explanation. This is in contrast to the constraints on agency statistics arising from the limited mandates of the agencies. For example, the geographic boundaries of attention, distinctions among persons (e.g., between adults and juveniles), and the weights of significance attached to data can be adjusted in any direction independently of varying legal or administrative definitions which restrict agency attention. The very definition of "crime" can be made as stringently dependent on or as independent of any criterion desired. (Thus, for example, if our interest is in economic costs, we may well wish to ignore for a given analysis whether damage was caused by an adult or a juvenile; if it is behavioral, we may not wish to limit ourselves by ruling out dangerous drugs that are not legally defined as narcotics or new drugs that may not yet have come under any form of legal control.)

6. Visibility of data costs. Special purpose data collection, despite economies of rationalized design and sampling, is likely to appear considerably more costly than using agency by-products as the original data source. In part, at least, these differences are illusory. Since statistical activities in the ad hoc endeavor are organized independently, all or most of the effort going into the system is separately budgeted and paid for, rather than much of it being absorbed by

administrative and operational agencies. (The time of basic informants in both agency and special-purpose statistical systems usually is not compensated by the statistical system.)

Low Operational Utility of Survey Data

The foregoing characteristics of the surveys as modes of developing knowledge of crime give them great value for some uses and users, but detract from their usefulness for others. Thus, for example, the survey method would be of very limited usefulness as a source of operationally useful information for a police department. It is an extremely costly and inefficient device for developing information on crimes--even on those crimes that victimize the individual citizen. In the highest crime-rate area in which we conducted interviews, if one asked respondents to give information on all criminal incidents of which they had been a victim in the preceding twelve months, the number of crimes for which the survey would yield information would be slightly fewer than the number of respondents in the sample. This is the case even though the victimization rate (0.83) determined by the citizen survey in this area proved to be vastly higher than that estimated from police offense data. An expensive survey of 1,000 cases covering a one-year period, therefore, yields fewer crime reports in a city like Washington, D.C. than the police register in a single week. It would be prohibitively expensive to use the survey method to develop information with the detail by time and place needed to afford concrete operational guidance for any police department.

To produce data sufficient for any analysis whatsoever, survey interviews must ask people about a considerable time-span. The further the respondent must stretch his memory, the greater the effects of forgetting and distortions of recall. The magnitudes of these effects are familiar in all surveys for objective data and they become particularly serious when the survey deals with ephemeral events, such as crimes, which yield incidence statistics, rather than with durable states of the subject suitable for prevalence statistics.

It was also found that most offenses that victimize citizens are not among the most significant of life events, and hence are not readily recalled in an interview. Relative to, say, births and deaths in one's family, major surgery, getting a job or losing one, or buying a new car, the bulk of crimes involve trivial and brief consequences for the victim. As a consequence, methodological analyses of the data from the crime interviewing survey showed pronounced signs of recency and telescoping effects, and other interview error.

Crimes that do have major consequences for the victim's person or property--and consequently those which deserve priority in police action--are so infrequent as to require huge samples to produce N's of analyzable magnitude. To illustrate the order of magnitude involved, if we use the Uniform Crime Report Crime Index as a measure of incidence, the expected number of mentions of crimes involving serious violence or threat of violence in a random sample covering 10,000 citizens would only be about 25. Even though surveys can yield some multiple of the crime rate that figures in annual police reports, the absolute magnitudes of the data from surveys with a feasible sample size are therefore much too small to permit of the kinds of analysis that would be useful as operational intelligence for law enforcement.

The guidance of police operations, however, is not the sole nor necessarily the major use of crime statistics. At the moment, their impact on public understanding and attitudes toward problems of crime and their significance for the development of intelligent social policies relevant to public order and safety seem of greater significance. These public knowledge and policy functions of statistics--which it has lately become fashionable to designate as "social indicator" uses--need to be distinguished clearly from the very specific forms of statistics useful at the operational level of criminal justice agencies as well as from the mid-range kinds of statistical data that are useful at the level of agency administration.

At the present time, the three kinds of uses are confounded mentally and organizationally with much mischief to our crime statistics. The same confusion, however, pervades much of the rest of our national statistical apparatus. Each of the three kinds of functions of information, I believe, is best met by a different form of organization and a different organizational position vis a vis other governmental activities. Such organizational separation can help serve the following ends:

1. To avoid confounding the purposes of knowledge with those of immediate action to the detriment of all types of use.

2. To separate those forms of information collection, aggregation and use which present grave hazards to freedom and privacy from those in which such hazards are minimal.

3. To extend the scope of available information to the entire range of pertinent phenomena, not merely those which are in a given agency's operational purview.

4. To place each type of informational activity in responsive relationship to its major users.

Organizational Location of Informational Activity

The field of criminological statistics has been handicapped because it has been centered organizationally at the middle one of three levels of use I have mentioned--that of agency administration. However, the great demand for statistical information on crime at both higher and lower levels of specificity has distracted the work at the administrative level, so that it has not met the purpose of any of these levels very well.

The lowest level, the most specific, of information needs is that for operational intelligence. This is the kind of information usable at the actual line operational units of an activity--by the police precinct, the individual patrolman, the parole officer or the judge. Such information tends to be useful in direct proportion to its fit to the organization and operating procedures of that particular unit, and to the very specific kinds of cases with which that unit has to deal. A national uniform crime reporting system inevitably imposes a set of categories that accords poorly with the very specific needs of any given jurisdiction. Although our statistical system places great burdens for recording and processing information for statistics on these lowest operating levels--on the cop, on the beat or the parole officer--I believe this level gets next to nothing in return. They give, but do not receive; I suspect as a consequence, they do not give very well.

To be of use at higher levels of organization--administration or management--information from the operational level has to be reordered, stripped of its idiosyncratic features, generalized. Our criminal justice statistics currently are products of the administrative agencies in the field. To the extent that there is a coherent rationale for the concepts, units of measurement, modes of aggregation, semantics of table labelling and the rest of the apparatus of these series, the rationale of the agency administrative perspective toward these problems largely obtains. Nonetheless, I do not believe that most of the current statistics are of any great usefulness as working tools of administration in the crime field. This is because of powerful factors that lead to their being prepared much more with an eye toward their external consumption rather than internal consumption by administrators and managers. With regard to the Uniform Crime Reports, for example, I can see readily how these series might affect the relations of a chief of police with his mayor and the public and, hence, his policy decisions. But I have difficulty in seeing much other intelligent impact they might have on his decisions on how to administer his department.

The final level of use is that which is required for general public understanding of the problems and for informing legislatures and other general policy-makers. Professor Wilkins has discussed the kind of rates that would be particularly instructive for the public and general social policy.

I believe there is an important role that can be played by initiating a statistical series in the crime field based on periodic surveys of the public and of special population groups, such as businessmen and institutional officials. Such a series can have primarily social-indicator uses, rather than administrative or operational ones. The potential uses of such surveys will be ideally served if they are sponsored or undertaken by an agency that does not have law enforcement as a primary responsibility.

Surveys of Operational Agencies

The foregoing arguments are not intended to gainsay the potentialities of agency statistics and the importance of improving them. This is particularly important for the administrative and operational uses of data, but data from agency sources can also contribute far more to social indicator knowledge. For example, many of the limitations of police statistics for which use of the survey method has been recommended as a corrective are not inherent in the law enforcement informational systems. Police agencies collect far more information than they process statistically and vastly more than they publish; for example, information on victims, characteristics of persons arrested, details regarding the nature and circumstances surrounding reported offenses, etc.

Such information is not readily assimilable into any broad statistical reporting system, however, because of the highly fractionated organization of law enforcement in our highly federal system. The limitations of resources, the absence of uniform definitions and practices, and the disincentives against full and accurate reporting by local jurisdictions and individuals can never be satisfactorily overcome by a voluntary and unrecompensed reporting system. To tap these resources will require data-collection systems similar to some which are used to collect information from samples of the public and of business enterprises. Such systems would make use of specially drawn samples of jurisdictions, as well as samples of personnel in each for certain specific data. In addition to voluntarily submitted questionnaires, trained statistical personnel are required to interview, observe, or code records as close to the original significant event as possible. Just as in the case of data developed from the public, the identification of individuals and individual jurisdictions may be kept confidential where these data are collected for social indicator uses. This is consistent with having this information serve the legitimate national purposes--knowledge and understanding--while safeguarding against illegitimate centralized interference in local affairs.

To give an illustration from the examination of unreported crime, it was assumed that the problem involved primarily the failure of citizens to report crimes to the police. Our results from a citizen survey in one city cast doubt on this. More offenses involving the victimization of private citizens apparently are not reflected in police statistics because the police do not count many citizens' complaints than because of citizen failure to report offenses to the police. I am sure that we would greatly reduce the "dark figure" of crime statistically if we were able to analyze time samplings of all incoming telephone calls for a national sample of police jurisdictions. (I hope statisticians will always bear in mind that an operator frequently has something he rightly feels is more important to do with his time than to fill out a form--even a "required" form--however.) An error of the past, however, has been to impose on the operational system an informational system that is heavily influenced by academic and political needs for data.

By the same token, much of the potential of surveys of the general public for measuring crime was not realized in the first trials of the method that were sponsored by the Crime Commission. This came about because in the collection and analysis, a dominant consideration was for the surveys to yield data that would be as comparable as possible to the offense statistics of the Uniform Crime Report and of local jurisdictions. As we have pointed out elsewhere (Biderman, 1967; Biderman and Reiss, 1967), the categories, units, rate-calculations, and, indeed, some of the underlying perspectives that are appropriate in looking at crime from a law enforcement perspective, are different from those which would be ideal for citizen surveys. Collection procedures must be adapted to the manner in which citizens experience crime. The major usefulness of the data thus collected also depends on using those analytic means which will reveal the crime experience of the citizenry.

References

1. Biderman, Albert D., "Surveys of population samples for estimating crime incidence," The Annals of the American Academy of Political and Social Science, 374 (1967) 16-33.

2. Biderman, Albert D., Johnson, Louise A., McIntyre, Jennie, and Weir, Adrianne W., Report on a Pilot Study in the District of Columbia on Victimization and Attitudes toward Law Enforcement, U.S. President's Commission of Law Enforcement and Administration of Justice Field Survey I, Washington, D.C.: U.S. Government Printing Office, 1967.

3. Biderman, Albert D. and Reiss, Albert J., Jr., "On exploring the 'dark figure' of crime," The Annals of the American Academy of Political and Social Science, 374 (1967), 1-15.

4. Ennis, Philip H., Criminal Victimization in the United States: A report of a national survey, U.S. President's Commission on Law Enforcement and Administration of Justice Field Survey III, Washington, D.C.: U.S. Government Printing Office, 1967.

5. Reiss, Albert J., Jr. (ed.), Studies in Crime and Law Enforcement in Major Metropolitan Areas, Vol. I, U.S. President's Commission on Law Enforcement and Administration of Justice Field Survey III, Washington, D.C.: U.S. Government Printing Office, 1967.

TRUTH AND CONSEQUENCES IN CRIMINAL STATISTICS

John P. Conrad, Bureau of Prisons

In this year of anxiety, our need for an adequate system of crime statistics should be obvious. Less obvious are the reasons why we do not possess such a system. The learned and the unlearned incessantly dispute the causes of crime and its remedies. They agree only in lamenting the lack of information on which to ground their contentions. From an abundant literature of wishful guesswork, orators derive widely diverging proposals for the cure of our national malady, but there is no serious dissent about the need for more statistics than we have. This concern becomes an obsession to those of us who are occupationally engaged in coping with the prevention of crime and the control of the criminal. It is reasonable, then, to ask: If the requirements are so urgent, what on earth accounts for the delay in meeting them?

The reasons are numerous enough to fill this paper and more like it, but they can be condensed into a fairly short answer. The elements of the answer would be lack of money, lack of consensus on plans for doing what must be done, and lack of conviction about the necessity for doing it on the part of the thousands of people who must collect and transmit the data to be transformed into statistics.

I shall now elaborate on the scope of what must be done. My version of the assignment ahead emerges from the perspective of a working administrator of one of the detached fragments of the system to be ordered out of a flow of loose data. I shall describe the present, propose the future and suggest some general principles. What I have to say does not necessarily represent official policy of the Department of Justice. It is rather a progress report on our preparations for a new day in criminal statistics, a new day made possible by the passage of the Omnibus Crime Control Bill, Public Law 90-351. We are at a stage when the formulation of ideas and the determination of priorities must be aired in such forums as this so that policies based on good statistical practice can be projected. What is at stake is the design of a system which will capture information to enable policy-makers and administrators to act rationally in dealing with man at his most irrational.

Now this is a country in which money is thought to be the answer to most questions of this kind. A recent estimate informs us that the nation spends about $178 million a year in the collection of statistics, mostly economic statistics. Of that total, about $800,000 goes tc criminal statistics. We might infer from this skewed allocation of resources that nothing is wrong which the multiplication of our present investment by a factor of ten or twenty or so would not correct. The deficiencies of our statistical system would be remedied and we could congratulate ourselves on a display of wisdom far exceeding the pound-foolishness of our predecessors.

I will contend that whatever problems money alone will solve, the creation of a responsive system of criminal statistics is not one of them. The quality of a statistical system is shaped by those who need it. Administrators who depend on statistics will see to it that the systems serving them will be accurate and informative to the degree that accuracy and information determine the success of their enterprises. It is true that aimless and shabby statistical systems will be tenaciously maintained by the bureaucracies which depend on their maintenance for their own survival. Flourishing systems can be identified by the uses made of them.

A classic example of a flourishing system is Uniform Crime Reports, published annually since 1930 by the Federal Bureau of Investigation. For most of the years since its inception, this collection has been the nation's primary indicator of criminal activity. Criticism of Uniform Crime Reports has accumulated into a considerable literature. I shall not add to it here. The achievement of the FBI in developing, maintaining, and improving a voluntary system of statistical reporting from several thousand police agencies representing virtually complete national coverage is formidable testimony to the administrative resources of that agency. It is an achievement which should be admired for what it is rather than deplored for not being what it is not intended to be.

First and foremost, it is a statement of work-load. It tells us how many crimes were reported to the police and what kinds of crime they were. It tells us how many of these reports were cleared by arrests. It relates these data to the annual population estimates of the Bureau of the Census, thereby creating crime index rates. It distributes the national crime load by state and city, indicates the percentages of clearances, and tells us about the size of each police department in the land. The tale is always lugubrious. There is always an annual increment in both the volume and the rate of crime, and the increment is usually of alarming proportions. We can draw our own conclusions but the implication springs out at the reader that the police need more or better resources, or both. Uniform Crime Reports should not be belittled as a social indicator. Its annual admonition is unwelcome. We may be unwilling or unable to decide on remedies for the conditions which it describes, and indeed this publication does not provide data by which options except more and better police can be chosen. But the part of the truth which Uniform Crime Reports conveys is ignored at our peril. We must set out to discover as much of the rest of the truth as we can. The concepts and the tools for the discovery of the truth are available as never before. As a people in serious trouble, we owe it to ourselves to get to work with the instruments for finding ways out of our difficulties.

If I have been respectful of the achievements of the FBI in quantifying the police work-load, it is because I am engaged in a parallel, though much more modest line of activity. For the last year, I have been responsible for the supervision of a similar but much less famous series, National Prisoner Statistics. If Uniform Crime Reports flourishes because it is responsive to the needs of police administrators, then National Prisoner Statistics shows how a system can survive even though nobody seems to know what to do with its product. Let me briefly recount its history and describe its problems.

The first effort to make a national count of prisoners was an undertaking by the Bureau of the Census in 1904. It was followed by reports in 1910 and 1923. An annual series began in 1926, covering all prisoners in Federal and State institutions for felons. This series ended in 1946 on the recommendation of the Bureau of the Budget that this function be transferred to some other Federal agency. With some logic the burden eventually fell on the Bureau of Prisons, which had never before seen itself as a data collecting and processing agency. A two-man unit was set up to design reporting forms, to induce states to use them, to receive reports on admissions, populations and releases, and eventually to produce annual reports of prison populations. Since 1950, when the Bureau of Prisons assumed this responsibility, its objective has been to attain complete national coverage for a head count of Federal and State prison populations. Lacking a field service, lacking even enough clerical staff to keep up with a timely publication schedule, we have been more concerned with our present burdens than with questions concerning the relevance of our service.

Our present publication commitments provide for an annual demographic report of all Federal and State prison populations. This publication is now four years in arrears. The last issue, published in the fall of 1967, covered the year 1964. Although it was handsomely designed and contained a completely detailed tabulation of all the data available to us, it was received without a ripple of acclaim or deprecation. Much more attention is given to our annual report on executions, which not only presents the data on the extremely infrequent executions which now occur in this country, but also charts the movement of population into and out of the nation's condemned rows. We have just published our report on capital punishment in 1966 and will get 1967 out of the way before the end of the year.

We also publish an annual summary of prison populations, which provides a patient world with data on total head counts. Next month we will publish our summary bulletin for 1966. I am optimistic about the prospects for publication of the 1967 summary before the end of the year. We are about a year behind the FBI in our publication schedule, and considerably farther behind in service to our field.

What accounts for this dismal situation? Our difficulties are only partly attributable to the small staff assigned to an intricate task. I will not deviate from the bureaucratic norm: we do indeed require more staff. But even if we were plentifully supplied with the analysts, field representatives, programmers and clerks which a statistical empire should rightfully enjoy, we would have trouble maintaining a timely schedule.

The reason is to be found in the traditions of the American prison. Wardens have always insisted on counts. A discrepancy of one calls for a recount, a process which will be repeated until the discrepancy is resolved. Wardens can be relied on to count accurately, but they have seldom thought it necessary to work out frequency distributions of their populations according to any base except housing and work assignments. Not long ago, I called the warden of a prison in a large industrial state to get information on the average age of inmates and the distribution as to length of sentence for certain offense categories. All this information was simply unavailable. It could not be obtained without initiating a special project which I did not feel justified in requesting. This situation is fairly common. Some states possess statistical systems which, I blush to say, are much more informative than the system we maintain in the Bureau of Prisons. The majority are hard pressed to meet the simple requirements imposed by National Prisoner Statistics. Age, race, offense, length of sentence, and termer status would seem to be simple information elements to provide for, but in the daily routine of most prisons their availability is not an urgent requirement. For a national statistical collection which should make orderly connections with other criminal statistics systems, this elementary demography is only the beginning.

Reflection on the problems of our ailing system convinced us that its condition was symptomatic of irrelevance. Even though some prison administrators must strain to respond to our simple questions, the answers do not constitute a body of anxiously awaited knowledge. Something can be made of our study of average length of incarceration for major offense groups, but even here the scholar is handicapped by a lack of comparability among the state penal codes. The rest of the data so painfully gathered ties into no other data except last year's and next year's enumerations.

What can be done? The answer can only be found if we put the prison in the perspective of a system of corrections. This is a system of probation, parole, jails, and workhouses which must be inter-related if the prison itself is to be economically and humanely used in the administration of justice. We can harry statisticians and administrators into making more precise and more timely counts according to prescribed demographic distributions, but if such counts cannot be given significance in terms of the effectiveness of the whole system, their reliability or the lack of it is beside the point. We must view the correctional apparatus of a state, or of the country as a whole, for that matter, as an entity to be understood in terms of its parts. Without this kind of understanding, we shall never know when or whether we are using any of the sub-systems of the correctional apparatus effectively.

A huge task can now be discerned. Not only must we know more and different kinds of information about our prisons, but we must collect similar kinds of information about our parole system, with which it is linked in an uneasy symbiosis. We must also inform ourselves about the use and effectiveness of probation, local confinement, fines and suspended sentences. If the flow of humanity from the courts into the various channels of correction can be accurately charted, we shall eventually arrive at some empirically supported principles to govern the disposition of the various kinds of people who fall afoul of the law. No judge can now say with assurance which dose of what disposition will best protect the public by the correction of any given offender. Worse still, there is no empirical base from which reliable advice can be given him. The administration of justice is a realm in which a priori doctrine necessarily reigns supreme. It is an uneasy realm. Subterranean rumbles warn us that the grounds on which we are standing, the assumptions with which we have always worked, are shifting beneath us. Change is ahead, and we need new kinds of data to prepare for it.

During this last year, our staff has been trying to project the basic requirements of a statistical system which would be a source of information about change, rather than the dutiful counting of heads. Taking as much and as good advice as we could get, we arrived at a plan of priorities which, we think, will justify the investment of the resources needed to carry it out.

First, we must get our own house in order. Our publications must be in the hands of administrators at times when they can be of immediate use. We will simplify our requirements for the head-counting process down to the lowest acceptable common denominator, one which will be a standard to which the least statistically competent state can repair. If there are states which cannot meet such a standard, we will omit them until they can. We will recognize the limited value such bulletins will have for administrators who need data of increasing complexity so as to assess the effectiveness of their programs. We therefore intend to develop a cycle of reports which will provide the field with data on institutional program participation, on personnel distribution, on the characteristics of prisoners, and on parole outcome. Obviously many states will be unable to contribute to such studies. All will be invited to do what they can. In some cases that will be a great deal.

I cannot estimate how long it will be before this process is complete. A creative staff with sufficient clerical support will be hard-pressed to accomplish all that needs to be done in the years immediately ahead. Without assurance of budgetary support, we do not know when we can even start such an ambitious program. Nevertheless, we are not without hope that we will have the wherewithal to undertake some development during the coming year.

While struggling with the modernization of this simple system so that it will serve the minimum administrative needs of prison management, we can contemplate in a spirit of admiring frustration the maturation of Uniform Parole Statistics. This is a data collection of parole case outcomes conceived and resourcefully maintained by the Research Division of the National Council on Crime and Delinquency. Supported by a grant from the National Institute of Mental Health, this statistical program aims at the provision of feedback to correctional administrators on the consequences of correctional decisions. Convinced that the improvement of parole decision-making depends on the collection and analysis of statistics of success and failure, the designers of this program worked with parole administrators for many months to identify the variables significant to decision makers. Once management was drawn into the system, the parole agency statisticians could be induced to design the codes and develop the machinery for a plan for random sample studies of parole release cohorts. Now, after some four years of hard work a series of impressive reports is becoming available. It is neither a comprehensive nation-wide system nor is it the last word in feedback to administrators. Few correctional managers are yet able to make full use of the kinds of analysis which Uniform Parole Statistics makes available for the first time. What I must stress here is the importance of this project in the development of statistical competent management. Just as good accounting practice depends on reliable bookkeeping procedures, good administrative practice must be grounded on some basis for estimating the consequences of decision options. The unavailability of any basis other than the ex cathedra pronouncements of the most articulate expert has gravely hampered the whole system of justice in arriving at a humane and economical foundation for policy and program. For the first time in our history the statistical machinery is beginning to be available for an administrative style which is second nature to industrial management.

Now here is a project which leads straight to the rationalization of correctional practice. If it is used intelligently it will revolutionize the processes involved in sentencing the offender. It should be indispensable. As matters now stand, there is some chance that it will be a vulnerable and short-lived orphan. It is dependent on the bounty of the National Institute of Mental Health and the continuing interest of the National Council on Crime and Delinquency. Neither agency has an enduring charge to support such a statistical system, nor is it reasonable that either should provide the resources to continue maintenance now that feasibility has been established. If the Bureau of Prisons had the resources, the National Prisoner Statistics program could logically assume responsibility for maintenance of the system. The complementarity of our present system, assuming its development along the lines I have projected, and the Uniform Parole Reports, is obvious. Once such a merger were made, we could look ahead to the day when statistical ignorance will no longer excuse uncritical reliance on traditional wisdom. But such a merger cannot be made without commitments which are not yet in sight. I would be alarmed indeed, were it not that concern is becoming general and in the middle distance the resources of a national criminal justice statistics center seem likely to come to the rescue.

There is a certain symmetry which I like to regard as the hallmark of the systematic administrator, if only because I am addicted to symmetry in all things. I can contemplate with aesthetic pleasure a statistical system which methodically covers all persons committed to adult correctional institutions and follows them through to the completion of their paroles. Once we have such a system we shall understand where now we flounder. But I am also aware that this kind of collection would represent only a fragment of the correctional process.

The symmetry which we seek and which, I hope, the field will attain in the foreseeable future, will provide for the collection of data covering all the possible dispositions of the adult offender. As of today, nobody knows within uncomfortably large margins of error the number of people confined in American jails. For that matter, nobody knows for sure how many jails there are. Not only do we not know how many people occupy the cells, tanks, dorms and other accommodations of the American jail, but we are also unable to say how they got there, what kinds of people they are, or what happens to them after they return to the streets. Yet many decisions are made about these people in spite of the virtually complete state of our ignorance. We grant and deny bail, we fix short terms and long, we decide whether jail should be a condition of probation, without even the most gross knowledge of the consequences of these decisions which are so routine for the decider, so momentous for the party of the second part. Worse, we plan new jails, settle on the kinds of staff and their numbers, and determine programs without the means for serious study of the alternatives. Usually such plans are made without even finding out what the alternatives might be.

I am arguing for a difficult but necessary kind of statistical collection, one which periodically establishes the dimensions of the field and between such global studies surveys the effectiveness and the movement within representative elements of the local confinement universe. I concede that the expense of a frequent count of all jail inmates would daunt any but the most compulsive statistician. I will also concede the difficulty of finding a logic by which any sample of American jails could be consensually seen as representative. Within some such structure the truth could be found, or enough of it to understand and facilitate the control of a squalid and often dangerous element in the administration of justice.

Somewhat the same problem confronts us in the study of probation systems. Most correctional administrators are impressed with their experience of probation as a method of control and assistance for the convicted offender. What also impresses us is that we don't really know much. The studies which have been done regularly find that probationers complete the course with happy endings. But we don't really know who they are; we don't know what happens to them while they are on probation; we cannot know what kinds of people are most likely to succeed or most likely to fail on probation. We don't even know whether the professional work of the probation officer was crucial to the probationer's success or whether it was the leniency of the deterrent aspect of the experience in court, or some other factor which made the difference between law-abiding behavior and a return to crime. Without feedback, we cannot blame the courts for their conservatism in the use of probation. Until we have data, the judge who grants probation is making a calculated risk without equations by which the risk can be calculated.

If probation is to be developed to its ultimate value in the disposition of offenders, we will also need data which will enable us to experiment with potential improvements to the system as a whole. We can experiment to the limit of our ingenuity with parts of the system. What is lacking is a concerted drive throughout the field to gather data by which the costs and benefits of state-wide major policy changes can be estimated. It is one thing to test hypotheses about differential caseload treatment in a county rich enough to afford probation research. It is quite another to take the course on which the State of California has embarked, where a conscious attempt is being made to determine the benefits to the state from the diversion of offenders from prisons to probation supervision. I am the last to minimize the value of testing hypotheses leading to the improved treatment of offenders. Microscopic studies of this kind are needed in much greater numbers than the field of probation is now undertaking if probation is to develop the confidence in the value of its service which will justify its increased use. But macroscopic studies of probation, requiring system-wide data collection and analysis are urgently needed if all the trade-offs are to be defined and understood.

Fortunately the indefatigable Research Division of the National Council on Crime and Delinquency is at work on the problem of macroscopic probation statistics. The lessons learned in the compilation of Uniform Parole Reports are being applied to a much more complex problem. It will be fascinating to see where this resourceful group will end with the solution of the sampling problem it will face. The difficulty of the task is magnified by a factor of about sixty when compared with Uniform Parole Reports, there being about sixty times as many counties as there are states: in probation it's in the counties where the action is. Moreover, so far as we can judge, most counties are in even less satisfactory a state of preparation for the tasks of adapting to a national data collection system.

Innovation by the National Council on Crime and Delinquency in correctional data collection are welcome in this field in which innovation is so difficult. To persuade a public authority to risk resources and reputations on new ways of conducting old business is a formidable undertaking. The evidence is to be seen in the retarded development of criminal statistics in this country.

But it must be recognized at the same time that private organizations cannot reasonably be funded to maintain elaborate statistical systems for the collection and dissemination of data for public use. It is at this point that public coordination of public functions must be made possible.

If you have followed me so far, you have discerned a disorderly array of statistical requirements. Some are well met, as in the case of the FBI's Uniform Crime Reports, which establishes police work loads and relates loads to resources. Some requirements are met by obsolescent services like National Prisoner Statistics. Some requirements are the object of imaginative experimentations and development, like Uniform Parole Reports. Some requirements are not being met at all, most notably mail and probation statistics. Many agencies are involved, and a case can be made for continuing the dispersion of effort in the interest of profiting from the continuities of communication and specialized expertise in the fields with which we are concerned.

With the passage of the Omnibus Crime Control Bill coordination of correctional statistics seems to be in sight. Among the new services for which this statute provides is a National Center for Criminal Justice Statistics. Its organization is still unsettled, but some of its functions seem to be predictable. It will not initially attempt to conduct data collections of its own to supplant or take over existing systems. What the country needs now is an agency which can make the possible connections between systems, which can arrange for filling the huge gaps between some of them, and, through analysis, help us all to create a national program for effectively reducing the incidence of crime and public anxiety. The Bureau of Prisons intends to build a system of national offender statistics out of the fragments of many systems now in being. New sub-systems will have to be created to fill voids with which no one has yet dealt. But the system which we will maintain will provide only one of the many data collections which the National Center for Criminal Justice Statistics will use. Our program will thus contribute to the mature understanding of its problems which an active society must have.

This is the note on which I wish to close. In his most recent book, The Active Society, Professor Amitai Etzioni distinguishes among four kinds of social organization at the national or macroscopic level. The passive society represent the emergence of national control over an essentially primitive social order. The over-managed society is totalitarian control through coercion without consensus. In both these kinds of society doctrine and direction are stressed beyond the need for information. The drifting society requires the mobilization of consensus so that action can be taken to meet its problems. But in such a society conflicts will exist which obstruct consensus until action is required by a crisis. The active society combines planning with consensus. In Etzioni's sense of the word no nation qualifies as an active society. Western societies tend to drift from crisis to crisis. There are signs, I think, that we are learning to combine planning with consensus in the economic domain, but we are certainly drifting in planning to meet the crime problem effectively. We have now drifted into a sense of crisis. We may continue to drift, but I think it is reasonable to construe the desire of the American people for control of the crime problem as an example of the formation of exactly that kind of consensus that is basic to the creation of an active society. Control can be achieved for a while by action without information, but we can predict that such a course will probably lead to even more serious trouble. We can achieve enduring control only through adequate information. The Omnibus Crime Control Law provides the apparatus for data collection and analysis by which rational programs can be formulated and chosen. We are ending a long history of drift. The means for finding the truth and measuring the consequences of our decisions will soon be in our hands. Action depends on concern and intelligence, but the fact that information will be available in increasing abundance justifies the belief that it will be action based on informed consensus rather than doctrinaire coercion.

VII

INTERNATIONAL HEALTH SURVEYS

Chairman, JACK MABRY, University of Vermont

EFFECTING A SURVEY OF HOMES IN LATIN AMERICA

Charles Roberts, Graduate School of Business Administration
New York University

They undertook a survey of the question. They did not hire metallurgists to examine samples, nor engineers to visit the site of construction. They took a public poll. Ten thousand people, guaranteed to represent every existing kind of brain, were asked the question: "Would you ride on the John Galt Line?" The answer, overwhelmingly, was: "No, sir-ree!

Ayn Rand, Atlas Shrugged

SUMMARY

Anyone who tries to effect a survey of homes in Latin America will encounter many difficulties, including the deficiency of maps and census data, the lack of trained personnel, and unusual beliefs of the population.

For the organization which has limited finances and political power, this paper sets down rules and gives examples for obtaining the necessary elements and points out some unusual sources of survey inaccuracy and some pitfalls that threaten survey failure.

INTRODUCTION

This paper is concerned not with the mathematical details of a Latin American survey but treats the peripheral problems of survey work. In the author's experience, these problems are more important than the mathematical ones, which are straightforward. This paper is essentially (2) with its 17 maps and aerial photographs and 20 candid photographs deleted for lack of space. You can obtain a free copy by writing me at the address given in the list of references.

In general, I will be reporting my experiences in a clinical-nutritional survey of about 250 localities in which approximately 25,000 persons were given a complete medical and dental examination and had their diet studied. The field work was completed over a period of about two years and was carried out by the Guatemala based Institute of Nutrition of Central America and Panama (INCAP), where I was employed for nearly the entire duration. The work was essentially paid for by the U.S. Government through ICNND, which letters stood for Interdepartmental Committee on Nutrition in National Defense and which went around the world studying soldiers, who must be in good nutritional state to protect the world against communism. Later it was decided that this made the U.S.A. look like a warmongering nation and the "D" was changed from "Defense" to "Development", and civilians were studied also. Hence, the problem of goals may not have been clearly defined. More recently, the letters are OIR for Office of International Research.

Although I will usually not explain how I solved the problem, I will point out the various difficulties, with the view that forewarned is forearmed. If I had to do it over again, in many cases I might proceed differently.

The largest survey that has ever been finished in Central America was performed by Sterling Products, which makes a popular analgesic, Mejoral. They have visited well over 50,000 dwellings, mostly in urban places, and could give a great deal of information on how it is done. The U.S. Public Health Service has done a health survey in Colombia and other places, and the ICNND and OIR have also completed surveys in Latin America.

GENERAL FACTS

There are many general facts that I have learned about Latin America, which may seem childish detail, but which may be the difference between infection or success in the field and approximate or exact data. Four such details are as follows:

1. The first time I ever slept with a mosquito net, I did not know that mosquitos will "bite" through the net. Since I am rather tall, I pressed my feet against the net at the end of the bed. When I awoke, I discovered that the soles of my feet were covered with bites, which drove me wild for about a week. Although in high malaria areas I took antimalarial drugs, such a large quantity of bites could have disastrous results. Will Dr. Jorge Cabrera spend the rest of his life on a diet of boiled eggs and milk because of intestinal amoebas?

2. When a Latin says that he is 19-years old, this may mean that he is 19-years old now or it may mean that he is only 18. Many people, when they complete their birthday, then assert that they are one year older than the birthday just completed. This will clearly not matter with adult data, but would cause great inaccuracy if two-year olds are classified sometimes as three-year olds.

3. Many Central American Indians believe that they have only a certain quantity of blood all their lives and so

refuse to give enough for a hemoglobin test because they "may have an accident in the future and will need all they now have."

4. One will find women breast feeding three-year olds, long after they ceased to have milk, because they believe they can't get pregnant if they are doing so, thus performing a method of birth control. In any health survey, the correct lactating status is very important, and as a matter of fact, because of the usual way of asking a woman's lactating state in Spanish, I have often seen lactation incorrectly recorded.

FOUR HURDLES

I will now mention four possible stumbling blocks which will face any serious survey in Latin America. If you don't think these items are important, I point out that an AID sponsored survey of 10,000 homes in each Central American country, with the unlimited funds of the U.S.A. backing it up, has failed.

1. Politics. I was surprised to find that one office worker brought her baby to work and spent her day babysitting until I found that she was a relative of a high government official. I wondered why it was not possible to arrange a purchase of gasoline, but it was possible to obtain airline tickets until I found out who was an airline company owner.

We were more than a little worried when the country strongman told us that there were no nutritional problems in his country, since "we have the tallest building in Central America." Now Guatemala, which has the most serious nutritional problems of the isthmus, has the tallest building.

2. Thieves. One must be on continuous guard against thieves both because of the large number of them and because of the general inefficacy of the police. When a Mexican store attempted to rob me of thirty dollars, the police refused to come, saying that I must seek a lawyer and file suit. After I was robbed of some $800 in clothing in Guatemala, the police failed to investigate although a formal complaint was made at police headquarters. The survey vehicle was broken into and personal effects taken in El Salvador, twelve maps and the results of three towns (which had to be revisited) were lost in Nicaragua, and a camera disappeared in Guatemala, to mention a few cases.

When one is far from the principal city of the country, he will find that it is impossible to obtain replacements for stolen goods and equipment. An object left unattended will not be left unstolen. Everything that has a top (hood, trunk, boxes, etc.) should have a lock and you should chain your suitcases to your car body and to the bed or the sink in your hotel room. I am acquainted with an unfortunate OAS official who lost her purse and her (official) passport because her suitcase was stolen from her El Salvador hotel room.

3. Manliness. Manliness, or in Spanish machismo, is very important at all social and economic levels, and can work for you as well as against you. Try taking an attractive girl with you when you enter a potentially difficult office. They will fall all over themselves getting together what she (you) wants. Remember, however, that the person with the highest position is the most manly. The one who actually does the work may be the least manly of all.

In Alianza, Honduras we came across a six-year old boy who had broken a leg. Unfortunately, it was Thursday and the visiting day of the town physician was Wednesday. Because the family was very poor, they could not afford to carry the child the 50 miles to Choluteca and were planning to have him wait for the doctor's visiting day. We tried to convince the father to put the boy into the jeep and we would carry them to the hospital in Tegucigalpa. He refused, saying that he could not afford it and would not beg charity, and would not be convinced that the service would be without charge. We were unable to get through to him until one of the nurses in the jeep told him, in an impatient tone, "Oh, put him in the jeep with an older brother who is more valiant than his father." With this, the father immediately put the boy into the jeep, climbed in himself, and we carried them to the Hospital San Miguel.

4. Misunderstandings. During the initial interview of the families of Tecapán, El Salvador, I was accompanying a Public Health nurse in the visits. We arrived at one home where the lady of the house refused to come to the door, declaring that she was very busy with the preparation of lunch. This was rather surprising, but since it was almost noon, we decided to return after lunch. At about two in the afternoon, we returned, but could not convince her to come to the door. We had almost decided to give up and take an alternative family, when the woman realized that she had been speaking with a Public Health nurse in uniform. She immediately came to the door, invited us in, and apologized profusely for not having invited us in before, explaining that she had thought I was a missionary who had come to try to change her religion.

Since we examined and weighed what each family ate, rumors spread that we were communists and would take from the rich to give to the poor. Many believed that the blood samples were for the soldiers in Vietnam. Usually, all such misunderstandings can be avoided by means of a talk with the town mayor, which will be discussed later.

THE INTERVIEWER

In the survey of Central America and Panama we had as interviewers both men and women, office workers, laboratory technicians, nurses, teachers, nurse supervisors, secretaries, social workers, and inspectors of sanitary facilities. Taking everything into consideration, I would summarize that there is no substitute for ability. By far, competence in the person outshines the sex or the current occupation. For political reasons I was forced to make use of a few inept individuals as interviewers.

Sex. In general, it is best that the interviewers be women. In our survey they had to ask the pregnancy/lactation state. These questions are best made by women, since men usually experience embarrassment and cause embarrassment with such inquiries. Many times one has to pry because she has not thought about it before, she believes it is bad luck to admit pregnancy, or she would really prefer not being pregnant.

Some people assert that men should never enter a Latin American home when the husband is away, because when he returns to find that a male has visited, he will get mad and be against the project. This is not exactly true because I have entered hundreds of Latin homes, when usually the husband was away, without the slightest problem during or after the visit. First of all, a man is rarely alone with the wife since there are always children, not only of the home but also of the neighbors, who want to hear what is being said, or there is a friend or relative present. Secondly, I usually wore a white robe, in which case I was not a man but a white robe. The truth probably is that the lady of the house is delighted to have a male in her home and get away with it.

Different. Without exception the interviewer should be attractive or at least not unattractive. Very fat people are to be avoided. The clothes should be neat and the hair should be arranged. In Latin America, the people like to see individuals who are different. Although I interviewed alone very little and usually entered homes with an interviewer, people realized I was not a native. I could see that they were anxious to hear me say something to see if I could speak Spanish; I rarely disappointed them.

It has been suggested that when the interviewer is different, the person may have the tendency to lie, but I don't believe this. Any time it is suspected that incorrect information is being given, the interviewer can try asking one of the children, who don't yet know how to lie. When the person being interviewed sees that his answers may be reviewed by his children, he ceases to tell falsehoods.

Ability with maps. There are some people who cannot read the maps of the towns. A nurse and a secretary were completely unable to find the dwellings indicated for the survey. They regularly made errors of one or two blocks, rarely visited the correct place, and more than once made an error of over 10 blocks because of the inability to understand the difference between northeast and southeast. The sanitary inspectors were excellent with maps because they had a great deal of experience making plans of towns showing sanitary facilities. One can easily improvise a test for the home visitor by simply letting him go on his own in a town with the map. A written test probably could be made to test one's map reading ability.

Secrecy. Nurses have the tendency to be secretive. It is probably a result of their hospital work where they are afraid to tell the patient the whole story or any story since the doctor may get mad. On the other hand, teachers are accustomed to explaining and are able to distinguish those who are going to understand little (and should be told little) and those who are going to understand a lot (and should be told details of the survey). Social workers have the ability to explain and to convince difficult cases to participate.

Mistreatment. Some interviewers never learn that the people are doing both themselves and us a favor, but act haughty with the families. In Latin America it is often the habit to mistreat social inferiors (employees, servants, salespeople, street peddlers, etc.). One nurse insulted the lower class families and did not know what to say to the upper class ones. Once she screamed at an uneducated lady, "You just don't know how to answer questions." It is difficult to spot these people at first because they treat their "equals" with poise and grace.

Avoiding misunderstandings. A uniform should be worn. I suspect that the Tecapán, El Salvador incident could be typical. Many Latin American families are extremely bothered by the number of

Protestants who come to their doors preaching the gospel. In El Salvador I saw many little signs on homes saying, "We are Catholic", while in Costa Rica the signs say, "We are Catholic. We do not accept Protestant propaganda."

When the sanitary inspectors identified themselves as such, people were sometimes terrified that they were going to impose some tax or enforce the installation of some expensive sanitary facility, the latter especially when the dwelling was also commercial. In the U.S.A. the census taker can cause a similar qualm by asking who is the head of household because of state regulation of young parents. The problem is easily solved: don't say or ask too much.

Some interviewers told lies to gain family participation. Several promised free milk and medicines. One even promised a new health center. These lies not only come back to haunt you but also make you feel very sad. You must be on the lookout that randomness is maintained. Some interviewers seek out poor families because "they need it more", and because they cooperate more readily, thinking they will get something out of it.

You must be especially careful with border towns since almost everyone there is engaged in some form of smuggling. Many live there illegally because they crossed the border clandestinely. They must not think that you are going to give them away.

Since it was important that the family structure be identified correctly, interviewers should analyze various sample situations. Ask them how they would code a family in which there are two wives in the same home (which occurred) or how to code the situation in which the wife has become old and the husband has taken up with her daughter (which occurred).

Clever. Interviewers for city populations have to be clever since one finds educated people there who know how their names are spelled and can ask basic questions. The whole project can look very silly if the interviewer cannot answer satisfactorily. The statistician wants to be sure that the sample he selected is representative, not just those who are uneducated, and hopes for little nonparticipation, a measure of interviewer ability. Help prepare interviewers by trial questions like "Who is paying for all this?" or "Isn't this a communist survey?"

One very alert interviewer was told that the family had heard about INCAP. "INCAP likes to fatten up children to eat them." Although she was barred from entering the home by the husband, menacing with his machete, she actually gained the participation of the family, by quick-witted appeal to their "common" religion.

HEALTH, HANDBOOK, AND INSURANCE

It is important that all survey workers have the usual vaccinations, including smallpox, polio, tetanus, diphtheria, BCG, typhoid, and yellow fever. Helpful will be (1) and (3). They should also take antimalarial pills. A complete medical examination should be made both before and after the work in the field and should include urine, blood, and stool samples.

Don't let them travel without adequate insurance. Once, INCAP found itself in the position of having to pay several hundred dollars to repair a crashed jeep, medical expenses, and prospective plastic surgery costs, because it had hesitated to get proper insurance. Of course, having insurance does not mean the insurance company will pay, and I speak from experience.

A handbook should be prepared that will include all important aspects of the work involved. It should include at least two completed examples of all the forms they will have to fill out as well as a sample explanation for the people visited. A complete set of codes to be used must be included. Explain why codes are necessary for the computer analysis of the data. Enclose a list of everything needed for work in the field plus a list of personal items they should have, like soap, toilet paper, etc. One person was worried on an air flight when the emergency exit said, "Cut along dotted line", and she didn't have a knife. They should carry their own water or at least carry water purification tablets.

Many people, including natives, are surprised to find that several places in Central America are cold, because of their altitudes. Water often freezes in Totonicapán, Guatemala. In Cerro Punta, Panama temperatures are so low that it was difficult to perform the clinical examination, since people had on undershirts, shirts, sweaters, and jackets, and did not want to take them off. Field workers should take adequate clothing.

Since we arrived in some towns to find that the map was of a different locality or did not even slightly resemble the place, it is necessary to explain how to make and prepare a map. You must include a table of random numbers and carefully explain it. Glue a sheet of important codes to the back of each worker's clipboard. Give him identification and, if possible, some kind of letter from an important government official. If he can carry a newspaper article about the survey, it will help greatly, even if people can't read it.

FIRST VISIT TO THE TOWN

The importance of the first visit cannot be too heavily stressed. When I worked in the National Institute of Arthritis and Metabolic Diseases, a survey was planned of a small town in Kentucky. The people were told that the doctors were coming, but it was not adequately explained that only their necks would be examined for the presence or absence of goiter. Someone maliciously passed the word that all the women were going to receive pelvic examinations. When the physicians arrived, there was not a woman to be found. It was a complete failure. Later the town was revisited, but not until the town physician had visited each family and made the situation clear. The second trip was a success, because of all the precautions taken and explanations made.

If this can happen in the United States, the situation is even more sensitive in Latin America, where education levels are quite low. Occurrences similar to the Kentucky episode befell us in Santiago de la Frontera, El Salvador and elsewhere. You have to have someone in the town on your side, someone who is respected, who is fairly well educated, and whom the people of the town can find when they have a question about what will be done. Usually the town has a mayor and this is the person most suited for the explanation of what will be done in his town. If he is not in agreement, the survey runs a considerable risk of failure since he has the power to convince all the people to disappear on the day of the examination and he can convince them to lie to the dietitians and anthropologists. I have yet to see a mayor that was not at least passively in agreement, and usually he is delighted that he is the person selected to be the spokesman for the project.

THE TALK WITH THE TOWN MAYOR

Almost all of the towns have a mayor (alcalde) although in Costa Rica he is the political chief (jefe político) and in Panama he is the corregidor. In Teustepe, Nicaragua the mayor lived in another town and did not have very regular office hours; in fact he did not come to town the day that I visited there. In Teustepe, we talked with the judge (juez) who was quite agreeable and gave us all the information we asked for, and, I am sure, was capable of helping us gain the confidence of the town. Other possibilities are the town doctor or nurse, the school teacher, or the richest looking person around.

I made a form for the visiting team to fill out at the original visit, a "Basic information of the locality" form. If you do not have such a form, you will find that many questions are not asked and the interview with the mayor or town leader is nothing more than an exchange of hot air.

THE BASIC INFORMATION OF THE LOCALITY FORM

This form's basic value is that it is orderly. Included on it are places for the interviewers to put their names and the person filling out the form has to put his name. This also is the case on each of the other forms. It makes a tremendous difference when the person has to sign his name because then he, and only he, is responsible for its completeness and accuracy, and he is being reminded of it. In one of the countries this system was not used and the consequences were a hundred almost worthless and costly to correct sheets. They should complete every form before leaving the home.

The time necessary to get there and the state of the road are important items of the form. You will find the system of roads in Central America quite bad and it would have to be known if the clinical team with all its equipment could get there. Some of the blood samples had to get to the capital city in less than four hours so that in some cases an airplane was needed and information about airfields is used. There are many airfields which are used for spraying crops.

Fiesta days of the community and nearby communities must be noted. Almost all towns have a week in which they celebrate a Saint, and it is hopeless to visit the place during this period, because no one is at home. Pay attention to national holidays also.

ADDRESSES

With the exception of certain parts of the principal cities, it is a problem to assign effective addresses to dwellings. Streets have no names indicated, homes carry no numbers for identification, blocks are generally poorly defined, and because the roads are owned by the government, and hence are "free", people may construct their homes smack in the middle of the street.

Typical city addresses are "From the church San Antonio ½ block uptown", "From the funeral parlor La Corona 1 (block) toward the mountain ½ uptown", "Avenue 15 de Septiembre", and even the very explicit "8th Street #1-69 zone 1". In more rural places the addresses may be "Main Street", "Cemetery Road", or more simply, "Street", and in one town I found three streets called "Main Street" and four called "Cemetery Road".

In many places the term "vara" is used to measure distances. A vara is equivalent to 33 inches, a more practical length of stride than a yard, although persons rarely know the exact measure of this quantity. In Costa Rica a block is considered to measure 100 varas even if it actually measures 50 varas or 200 varas. Hence, we have the addresses "450 varas to the north of the store The Last Drink (450 varas al norte de la pulpería La Ultima Copa)" and "From the bakery La Concepción 325 varas to the north and 150 to the west." Don't let yourself be fooled by the habit of giving distances in time because a place five minutes away may be five minutes in car, on horseback, or on foot.

You won't get too far asking to what address they have their mail sent. One person replied after thinking some five seconds, "Well, I really don't receive mail." Another said, "Post office box 29." Another told me, "Mr. Juan Villaverde, Teustepe, Nicaragua. Everybody knows me here."

Actually, if one has to put an address on each family, the task is not impossible. Every town has a plaza and you should, as a rule, ignore the existing names of streets, since everyone else does, and call byways from north to south Streets and those from east to west Avenues. The plaza is bordered by First and Second Streets, and First and Second Avenues with Avenues 1, 3, 5, 7, etc. to the north and 2, 4, 6, 8, etc. to the south. It is thus possible to give an address as "Street 4 between Avenues 1 and 3." As a further identification, the malaria control number of the dwelling may be recorded, if it has not washed away. It may also happen that Public Works or the Department of Health has placed a useful number somewhere. In the city it is wise to place a number in each block as a further aid to giving addresses, like "Western side of block 3304."

FAMILIES WHICH DO NOT PARTICIPATE

Very few families did not want to participate in the clinical-nutritional survey. It is not possible to give very accurate statistics on the types of persons that don't collaborate since there are so few and because the reasons are usually a bit odd. In many of the towns all of the families that were asked to participate agreed to and often people stopped us on the street to ask us to include them or to ask why they were not invited.

One lady of a selected house did not want to participate because she was mad at us for interrupting her soap opera on the radio, La Esposa Infiel. One refused to cooperate because she had a "fear of communists". I remember a case in which the husband and wife were not getting along well and she did not want to participate because he wanted to. She exclaimed that she was planning on leaving at any minute because he had other women and that she would probably not be around when the clinical team came. One lady said that she didn't like doctors.

Although I don't present proof, I would say that, in general, the people who don't want to take part are those that live on the outskirts of the town and have a fear of what is going on inside of it. Of course, it could be argued that the people that live on the outskirts of town fall into two classes: those families that are very poor and hence cannot afford a better lot closer to the center, and those families that are among the richest and live on the outskirts because they want to be near their lands. These relatively rich have a great fear that you are going to trick them out of what they have while the very poor are often uneducated, ignorant, and unsure of themselves.

It is difficult to get men to come to the clinical examination because they have to work. The 75 cents that they will receive for a day's work is extremely important to the family. A survey should pay the men their salary to encourage them to come, especially if it is essential that men be examined. In Nicaragua it is common that if a person works from Monday through Saturday without missing a day, he is paid also for Sunday. Therefore, if he goes to the clinical examination on say Wednesday, he not only is not paid for Wednesday but also is not paid for Sunday. It would be a great sacrifice on his part to attend an examination and lose two days' pay for doing so. A letter to his employer will help.

It is a good idea to give the families something as an encouragement to participate. We made a Polaroid photograph of each family, which was appreciated tremendously. You would be amazed at the effect of a plastic bowl, powdered milk, or vitamins. In Honduras we were able to give away toothbrushes and toothpaste that had been presented to us by the manufacturer. Teeth were pulled free.

It has been suggested that in the United States the people who want to participate in health surveys are those in good health, who want to demonstrate that they are healthy. This does not apply to rural Latin America, since few are really healthy, although one would expect this to happen with Latin Americans who are fortunate enough to receive medical care.

In the city the reasons for refusing to participate are somewhat different. One well-off family in San Pedro Sula, Honduras said that they didn't need any medical examination since they "could afford to go to the United States if there were a medical problem." In the city one may have to appeal to the higher instincts or to the education of the individual. In Guatemala City a lady of the house said that they didn't want a medical examination because her husband was a physician. We called her husband at work and he came home immediately to talk with us. He was very agreeable and the entire family participated. There are a lot of "new rich" who, as the Spanish expression goes, "still have the stain of eating plantains" and retain or acquire certain antisocial ideas that flag their economic status. In any case, don't hesitate to call a person's physician, relative, or friend to ask aid in clarifying the requirements of your survey.

I have seen several families that believed that the clinical examination would be gratuitous but were afraid that afterwards they would be told that they were sick and should buy medicines, for which they had no money. Without exception, these people were eventually convinced to participate because, the truth is, they are very interested in their health.

The majority of families that don't collaborate are very poor and lacking in education. The extremely poor think that they are receiving a handout and because they are proud and don't want people to think and don't want to feel that they are begging, they refuse to participate. On the other hand, in Panama we found people who proudly boasted that they didn't need doctors because they visited a healer (curandero), when necessary.

AVAILABLE MAPS

The basic map is that of the Cartography Section of the Department of Statistics and Census. These maps have a good basis since they were made from aerial photographs, and at least the scale is correct. Many times the map will say, "Made 1965 from aerial photographs" and one is glad to have a map up to date, then reads further to find, "...taken in 1954". Sometimes due to flood or because the original photograph was in a bad state, the map has not been accurate in the field, but some 80% are correct as far as divisions of blocks are concerned. Because some of the aerial photographs of Buenos Aires, Nicaragua were unsatisfactory, half of the map was a neighboring town, Pueblo Nuevo, without pointing this out. One would not realize this until he went to the place.

These maps are usually available for all cabeceras, or county seats, and often for other localities that are of importance like Siuna, Nicaragua, which is a mining center. There is an original in the Department of Census and one must make a heliographic copy with special paper. The obtaining of copies may be traumatic and will be described later.

The other basic map is that of SNEM (National Service for the Eradication of Malaria). SNEM goes everywhere that a malaria carrying mosquito goes. These maps, in general, are quite good and exist for many more places than those of Census. The census map of Corn Island, Nicaragua showed only 79 dwellings while that of SNEM had 289. Use of the malaria control map saved at least two days' work in the field!

In Nicaragua, SNEM recognizes 5007 localities which, unfortunately, are only those which have a malaria problem. Many places, like Totonicapán, Guatemala, do not have a malaria problem because of their high altitudes, and hence their low temperatures. Others, like Cabo Gracias a Dios, Nicaragua, are at sea level, but do not have a malaria problem since a strong wind, originating from the sea, washes the town, and a mosquito cannot reach it, even crawling on all fours. In other cases, like the department of Darién, Panama, a malaria control worker was stopped by six naked Indians who asked, "Where are you going, and why?" When he responded he was sent to kill the mosquitos that cause malaria, he was told, "You may not pass", and he did not. There have been similar incidents in San Blas, Panama, but during some malaria emergencies, the military has accompanied SNEM.

It is a fact, but I offer no statistics on the subject, that the malaria people find about 10% more dwellings than the census people. A SNEM official told me that it was because the census was taken by irresponsible high school students, by the town drunk, or by the unemployable, and that Census simply did not visit all the homes. On the other hand, a Census official told me that SNEM is paid by the number of dwellings, since they go from house to house spraying, and naturally find more dwellings than there really are.

From my own personal experience, I would say that there is a good bit of truth in the both, but point out that SNEM has interest in all buildings while Census is mainly interested in homes. In warm climates the kitchen is often a separate building because it is hot, and the family is trying to avoid heat as much as possible. In this case, the census map may or may not have the kitchen structure shown, because it is generally small and may not show in aerial photographs; the census figures will not count the kitchen as a home, and SNEM will list it as a building. This double counting of homes has been a problem, but there is no way to avoid it except to visit the town.

There are maps of the Sanitary Inspectors of the Ministry of Public Health who have the responsibility of checking on the health standards of toilets and sewer facilities. The problem is that the map is invariably in the town or somewhere in the field, and you must go looking for it. You had better try to locate the inspector first with a telagram or some other way, because he has a wide territory to cover, in general. You can find maps in the Institute of Cartography (not Census), which has aerial photographs if you know how to use them, Public Works, Public Roads, Institute of Homes, the Municipality, and any organization that has made surveys in the country, like ROCAP, ICMRT, MARU, and INCAP.

There is considerable quality variation in the SNEM maps since they are made by people who have as first interest the killing of mosquitoes and are amateurs at map making. In one town the malaria map showed the cemetery in the south when it was actually in the north and showed the church on the east side when it was actually in the west. In another map, half was correctly made but, for the other half, north and south were interchanged with the result that the second part had to be read from the back side of the paper. In another town which did not have a census map, the selection of the homes for the survey was made in the office with the SNEM map, but when we reached the town, we could not locate the church or any home as indicated on the map and were forced to make another map on the spot. On the other hand, the SNEM maps have generally been revised several times and are at least as dependable as those of Census.

SELECTION OF THE TOWNS AND HOMES

The main problem is that there is a serious lack of appropriate maps. With the exception of Costa Rica, there are few maps of the rural areas that indicate homes, although the various Institutes of Cartography have numerous rural topographical maps. Furthermore, few of the towns have census maps except cabeceras. It is difficult to get population figures for noncabeceras since the figures are not generally published. In each of the countries there are about 250 cabeceras (e.g. 170 in Honduras) although there are approximately 9000 populated localities (e.g. 8595 in Panama). On the other hand, the census figures for cabeceras are often not very precise due to the fact that the boundary of the town is never accurately defined and often depends on opinion or, at other times, on whether or not the family has running water.

In some census maps, for instance that of San Rafael del Norte, Nicaragua, there is not a single home shown and, in this case, the census sectors had not been recorded, so that it was known how many dwellings there were in a sector, but not where the sector was. There is no malaria map of San Rafael del Norte.

In some towns, as was the case several times in El Salvador, there seemed to be only a block from the center of town to a certain house, but as it turned out, the home was in effect miles away because there was a mountain, a ravine, or jungle separating the town square and the dwelling. Locating homes in San Pedro Ayampuc, Guatemala, a mountain village built on and around several mountain peaks, would be a tremendous struggle.

If, in the United States, one wanted to know the number of houses in a community, he would drive to the place and count them. It would take him three days to arrive at San Juan del Norte, Nicaragua, which is a cabecera; he would travel first over the huge and dangerous Lake Nicaragua, and then by small boat would cross the shark infested river Río San Juan. He must take his camping equipment because there is not a home that can or will accept him. He must carry everything he needs to live and fight off disease. If he runs out of food, money cannot always buy it. There is often no telegraph, no type of airport, and it can be expected that any mail may be lost.

In general, it would be practically impossible to have a successful survey in which all localities of the country could have been in the sample, because of map and transportation reasons. Usually, a survey would include only cabeceras, with the hope that there is little difference between cabeceras and noncabeceras. While I allowed myself to be convinced of this at the beginning of the survey because of time and expense requirements, I later pressed for more verification of this point, and now am extremely skeptical. Our survey considered separately the large towns by studying the most industrial and rapidly growing city in each of the countries.

The preparation of the maps for work in the field was generally made by combining both the census and the malaria maps of the locality. In some blocks, SNEM found more dwellings than were indicated on the census map. In such case, the larger number is assumed correct, and the selected dwelling is, for example, "the third", where the starting point and direction are indicated. In the blocks in which there are the same number of dwellings on both maps, or more shown on the census map, the very home selected is circled. When a dwelling was vacant, destroyed, or otherwise, the interviewer was directed to visit the following home in the same block. One should prepare all he can in the office, because revision in the field is difficult, because of the lack of transient accommodations, and because it is desirable to be in the town as short a time as possible in order to avoid infections.

TO OBTAIN THE MAPS

The SNEM maps are the easiest to obtain because these people have great pride in what they do and are eager to gain more recognition. Almost all of the maps are size 8" x 10" and nearly all of the offices have a Thermofax machine for making copies. If you need several copies, you should carry your own Thermofax paper (about six cents a sheet) because there is no budget for giving it away.

For copies of census maps, you can expect to supply the heliographic paper and the ammonia that is used or have the work done by a private firm. Private firms charge about eight cents per square foot of copy, although some may run three times that much. Thus, a sizable job done by a commerical concern will be expensive. The paper comes in 50 yds x 42" rolls and, depending on the country or the store, costs from $10.50 to $25 for the type imported from the United States and from $7.75 to $15 for the type made in Costa Rica, if you can find it. There are ten-yard rolls to be found, but I never located them except in Guatemala and Costa Rica where they cost $4.50 and $3.00 for the United States and Costa Rican types, respectively. In Honduras I found only the American type although it was stated that they usually have both. The ammonia ranges from $1.75 a gallon to about four times that much, depending on the country and the store, and a gallon is enough to process about 100 yards of paper.

You can have serious problems in getting the census maps and must expect to pay at least one-half more in paper than it actually took to make the maps. Requiring excess paper is almost universal. In one country there are special procedures with the Director of the Census Section before maps can be made and when final approval is obtained, one must then try to arrange with his relative in the Cartography Section to have the maps made. In this case, I was told that it would be weeks before they could produce the copies, but if I paid, a person could make them after working hours. I paid and was amused that the maps were made during working hours.

I am ashamed of that little payoff, since if you have to bribe someone, you haven't done a good job. The next time I asked them for maps, there was no difficulty since I worked through the Director of the Department of Census. It is an excellent idea, if the work is fairly large, always to contact the Director, since, in general, he (or she, in Panama) is a person of competence and honesty. All communications with him should, of course, be only in Spanish.

One of the most difficult times I had in getting census maps was brought about because of a U.S. Census Bureau advisor. His situation was ludicrous. Although he had been in Central America over two years, he was unable to speak Spanish fluently, while he essentially directed an office of over 40 people during data tabulation. He directed the staff by telling his secretary in English what he wanted, and then she, in a condescending manner (since she was the only English-Spanish person in the department), directed the work. She, because she spoke only English to Americans, and a Mexican railway ticket seller who had overcharged me twenty pesos were the only two people I have encountered in Spanish speaking countries that did not "understand" my Spanish. This American official took an immediate dislike to me when he found that I had more in common with his employees than he did.

THE LARGE CITIES

The large cities may prove to be a serious problem since the number of dwellings by block is not usually known. However, census figures are always available for sectors, which have the size of some six or seven blocks, although there are certain exceptions and at times the figures are in error. There may be 300 or more homes in a sector, and you would find it almost impossible to locate the 125th family. You will discover comunas, where behind a single door live 20 or 30 families. Usually there are no nameplates so you must knock on the door to see how many families live there.

It is usually best to select a block inside the sector either at random or by the approximation that each block has the same number of dwellings. The dwelling could be chosen by one of the following three methods: first, one could go to the block chosen, count the number of dwellings, and randomly select one of them; second, one could divide the perimeter of the block into say 20 parts, select a number between 1 and 20, and go to the approximate location corresponding to the number selected; third, one could say that the block will have at least 10 dwellings, and select a number between 1 and 10, and go to this home or always select, for example, the fourth home in the block. The third method lacks some degree of randomness, but is easy to explain and satisfactory for many purposes.

One economic researcher selects a block at random, then visits the houses numbered 2, 5, 8, 11, ... until the block is exhausted, or the researcher. He then continues to select blocks at random until he has the total number of families desired. This is clearly not perfectly random, but suffices for him. The El Salvador Department of Census did a work force survey in San Salvador by a random selection of blocks and the random selection of a dwelling in the block. They did not take advantage of their own data, since they ignored the fact that some blocks have more dwellings than others.

SPECIAL GROUPS

There are two special groups in Central America and Panama which have no equivalences in the United States. These are the workers on plantations and police or police-military. For example, in Fincas de Sánchez, Costa Rica, one plantation had more than 75 dwellings which all belonged to the plantation owner. These housed a large group of people, and the families were of extraordinary size. Several of the dwellings were empty because they were not needed until the crops were to be harvested. It is best to visit to see how many homes are vacant, if possible.

The military are an interesting group and are very different from country to country. Since it is better to join the army than to starve, one finds 12-year old soldiers. In Costa Rica there are no soldiers and the police are people who have homes and live like civilians when not on duty. On the other hand, the police of Nicaragua are soldiers who live as soldiers. In Nicaragua and the other countires, except Costa Rica, there are soldiers in small numbers, like one or two, and up to ten in almost each town for purposes of keeping the peace. They generally live in the military quarters and if you want to study them, you must go to the government building that is both their office and their home. Completely random selection of soldiers is difficult, if not impossible, since for security reasons they refuse to divulge information as to the numbers and locations of troops.

KEEPING THE DATA CLEAN

When I learned that X rays were made to determine bone growth in children and bone density in adults, I asked if pregnant women were also included. I was told that women were always given a lead apron. In the field I found that this was not the case, and in fact no one was ever given the apron, although it was always carried. No one could tell me if the equipment was collimated, but a head physician (who had said that all women used the lead apron) informed me that it was a low level machine so didn't really matter. I'll bet if it were his pregnant wife being X-rayed, she would get a lead apron.

With this hint that I was being lied to, I was angry, and began looking for other operations that were not what they were supposed to be. I didn't have to look far. The medical duplicates (two physicians examine the same patient) were being done with collaboration, and at times either the original or the duplicate was changed to obtain uniformity. I forced a showdown to distinguish between duplicates to train and improve the quality of the data, and duplicates for the computation of standard errors.

For some time I tried to get the dentist to do duplicates with his assistant, but he could not see the need and would not be convinced to do so, partly because of the lack of time. After five days of my insistence, he did two duplicates, and handed them to me with a bored look. I skimmed the pages and, for one person, pointed out that he had found 6 more cavities and a gum disorder more than his assistant. He sought out the people, had a training session with them, and from then on duplicates were done without complaint, with a corresponding improvement in the data.

Since the U.S.A. was footing the bill, there were American physicians assigned to the work. Unfortunately, American physicians both are used to making diagnoses before there are any clinical signs and have never seen many nutritional disorders, like streaked fingernails. (Streaked fingernails should not be confused with white spots in fingernails, but appear as light and dark arcs.) After I had finished with my work at reception, I reviewed the clinical forms. One American physician was finding many type 1 goiters and streaked fingernails. I stood behind him, not for very long, and after he recorded goiter and streaked fingernails on a case, I examined the person (although I am not a physician, I have attended several training sessions and felt several goiters) and said that this was not a goiter and the person did not have streaked fingernails. A conference resulted, and I was proved correct, with the resulting improvement in the data.

DISASTERS

You must expect disasters and plan for their occurences. Work had begun in El Salvador when a serious earthquake, which destroyed numerous bridges and drastically upset the usual pattern of life, caused us to give up the project for several months and return to Guatemala. One driver died in an off-duty accident. Of the interviewers who made original contact with the families in the field, one broke a leg, a (very attractive) one received two four-inch facial scars, one caught chicken pox, one intestinal amoebas, and two were robbed. One physician dropped out for health reasons. An emergency aircraft landing was necessary. Floods occurred twice. On about 6 occasions, the town was reached by the interviewers but could not be reached by the medical team because of rising rivers or rain-damaged roads.

COMPUTER SUPPORT

If you are interested in computer support or require computer usage either to plan or to analyze your survey, you will be pleased to know that there are several computers in Central America. On the other hand, IBM of Guatemala is not IBM of the U.S.A. For example, when the University of San Carlos received Guatemala's first IBM 1620 computer, it did not occur to the company that the computer's typewriter would need paper, so none was available. The University engineer in charge of the project looked all over Guatemala for paper which would be suitable, since it was troublesome to keep inserting 8½" x 11" sheets. He tried rolls of paper towels, but they did not work well, although you could then dry your hands with old printout. Finally, he discovered that Tropical Radio has paper in rolls, tearing off a piece for each telegram. The last time I went to the University, he had inserted a Tropical Radio roll into the typewriter, backwards of course, and was managing fine, except that the roll was just a little short of being wide enough. Fortunately, many of his programs were written for Tropical Radio paper.

CONCLUDING REMARKS

The most enjoyable aspect about working in Latin America is that the people are very pleasant and willing to cooperate. There is always someone (wife, children, or old folks) in the home when it is visited except in some 3% of the cases. On the other hand, since men must work, you will not find an adult male until late afternoon or nightfall unless he is old, ill, or the husband is away. A Latin child of 10 is often more mature and responsible than an American child of 16, and both can and will give excellent answers to questions. You will not have to set aside an extra day to get people to cooperate or because a family was not at home on the first day.

The American statistician for a large health survey in Colombia once stated that his survey did not experience "any of the difficulties" of the clinical-nutritional survey in which I worked, and that he incurred "97% response". The naive individual would probably think the Colombia survey's statistician both very fortunate and very successful. To me, however, it points out that he just didn't know what was going on at the nitty gritty level. With some Central American Indians, which are culturally similar to Colombian Indians, there was only 40% response while an 80% response could be considered good for adult males. While I would not be impressed by 97% saying they would participate, a 97% response to a physical examination is an impossibility in Latin America.

Although the field work for the INCAP survey was completed in May 1967, there is as yet no published account of the results.

ACKNOWLEDGMENTS

This work was in no way supported or approved by the Institute of Nutrition of Central America and Panama (INCAP) or by any branch or office of the United States Federal Government. Although experiences narrated in this paper took place while the author was employed as a statistician with INCAP, the entire article was adapted from personal papers, letters, acquaintances, and memory, while he was employed at New York University.

The author wishes to thank his wife Edna, who accompanied him on many trying trips throughout Central America. For example, we awakened in Juticalpa, Honduras to find that a rat had disposed of the breakfast breads we had in our room and to observe that we were covered by bites from bedbugs. Although her left breast was bitten by bedbugs so many times that it turned blue, she maintained her good humor at being referred to as "the girl with the blue boob".

Special thanks to Berta Mendizábal, Carlos Martínez, Aníbal Avila, Carlos Sandoval, Emma Guadalupe Jacobo, Carlota López de Cardona, Mariano A. Pinto Magaña, Olga and Máxima, who couldn't do the work but taught me much, Orlando Danilo Sota, who, to mention one of many feats, managed to bring our Land Rover safely to a halt when a brake cylinder broke and left us without means of stopping. With only a small wrench and a nail head he was able to seal off the cylinder so that we could continue on our way, although the wheels pulled to

one side, when braking, to be sure. José Trinidad Fiallos, Roberto Hernán Rosales, Agosto Aguilar, Francisca Mejía Romero, Angela María David, Raúl Castillo Borge, Enrique Lanzas Balladares, Victor Pou, Julio Guzmán, Felipe Arellano, Gladys Pineda, Olivia Miranda, Mery Morales, Teresa Anglada, René Sánchez Bolanos, Ana Isabel Montoya Monge, Ana Estela Arias Flores, Luisa E. Quesada, Francisco Sibauste, Hernández Quirós, Paco Rivera, who risked his life destroying mosquitoes and making maps until he lost his job for political reasons after 12 years, Eddie Chen Solé, Indalecio Valdés, students of CENADAL, and many others. Que Dios los guarde a todos.

REFERENCES

(1) Most, Harry, Editor (1964). Health Hints for the Tropics. Fifth Edition. The American Society of Tropical Medicine and Hygiene.

(2) Roberts, Charles (1967). Effecting a Survey of Homes in Latin America, 45 pages, privately published booklet. Write the author, 100 Trinity Place, New York, N. Y. 10006, for a complimentary copy.

(3) U. S. Public Health Service (1963). Immunization Information for International Travel.

A COMPARISON OF METHODS OF THE COLOMBIAN AND U. S. NATIONAL HEALTH SURVEYS

Margaret D. West and Garrie J. Losee, Bureau of Health Manpower

The Colombian National Health Survey was conducted by the Colombian Ministry of Health and the Colombian Association of Medical Faculties between September 1965 and June 1966. This survey was designed to provide information on health characteristics of the population of Colombia, a developing country of 17 million people. Data collected from 8,800 household interviews and 5,000 clinical examinations are now being analyzed and published.

Since 1957 the U. S. National Center for Health Statistics has been conducting a continuing Health Interview Survey. Beginning in 1959 a series of clinical examination surveys or cycles, directed toward various age-segments of the United States, has been conducted as components of the Health Examination Survey. The first of these cycles is singled out for comparative purposes, because it provides the widest coverage of the U. S. population to date. For the purpose of our presentation we will refer to these two data collection activities together as the U. S. National Health Survey, although, in fact, the National Health Survey is a broader program including other data collection mechanisms. We do not intend this comparison of methods of the two national health surveys to be an evaluative one. The purpose, rather, is to describe deviations in methodology in the Colombian survey from what could be called the "standard" methodology of the U. S. National Health Survey. We and the other architects of the Colombian health survey would have been satisfied to have been able to duplicate the highly developed methods of the U. S. National Health Survey. It should be mentioned that all but one of the seven United States consultants to the Colombian Survey were at the time directly associated with the U.S. survey. The Colombian staff members of the survey observed the operations in the United States and consulted at length with their counterparts in the National Center for Health Statistics and the Bureau of the Census. It is therefore apparent that a definite effort was made to transfer knowledge and experience gained in the United States to Colombia.

However strong the desire for duplication may have been, it was apparent that modifications would have to be made in the "standard" methodology to accommodate both the special set of limiting conditions in Colombia and the different objectives. Major differences between the United States and Colombian health surveys are not many, nor did they occur by chance. We believe it is worthwhile to describe these differences, why they were felt to be necessary, and their implications in order to create a better understanding of the factors which can affect the design and operational procedures of a health survey, particularly one which is to be conducted in a developing country. In transferring, or adapting the U. S. methodology to Colombia we had to make adjustments to meet a number of special problems. At the same time there were advantages in the Colombian situation which added to the accomplishments and satisfactions of the undertaking.

Among the problems, five should have particular mention. The first of these was the limited experience and understanding of the concept of a probability sample. The only Colombian experience with sampling on a more than local basis of which we learned was the use of sampling techniques in agricultural studies. The 1964 National Census in Colombia was a 100% enumeration, with no sample for any aspect of the undertaking. In 1964 the National Statistics Department (DANE) considered requesting the assistance of a UN Expert to select a sample of the Census for advanced tabulation of the 1964 Population Census. Research in the social sciences was limited to non-probabilistic case studies. A second and related problem was the limited understanding of computer technology. This problem will be discussed later under the heading of data processing. Third, since the last Census had been taken in 1951, 13 years earlier, population counts, maps, and other sampling materials from that Census were considerably out of date. Special effort and methods were required to overcome these limitations in selecting an area sample. Fourth, the terrain of Colombia -- the several series of high mountain ranges which can be crossed only by airplane, the areas which can be reached only by incredible mountain roads, or only by boat or on horse back -- was a challenge both in the design and execution of the survey. Fifth, the high rate of illiteracy caused problems in the availability of interviewers and other staff. The illiteracy level also had to be taken into account in designing questionnaires and examination procedures. Added to these problems was the threat of political disorder. Concern had been expressed over the possible erruption of violence after the forthcoming 1966 national elections similar to that which followed the 1948 elections. Guerrilla warfare continuing from the 1948 disorders raised doubts about the feasibility of conducting a survey in some sections of the country.

Set against these problems were a number of real advantages. First was the sense of national purpose of the Colombians. This study had urgent meaning to them as a keystone of the development of a national health plan. It was undertaken with the active support of the President of the Republic and his Cabinet. The seven medical schools of Colombia, through their Association, took active part in the planning. They provided medical students as interviewers, and residents from teaching hospitals as clinical examiners. These young people, in turn, brought to the study a sense of discovery and interest in their own country which was evident as enthusiasm throughout the survey. This, together with the excellent staff of the Ministry of Health, made for high professional standards for the work. Together, these assets made also for remarkably fine community cooperation, and in turn for very high response rates both for the household and the clinical surveys.

OBJECTIVES

The programs of the U. S. National Health Survey were developed on the basis of recommendations made in 1953 by the Subcommittee on National Morbidity Survey of the National Committee on Vital and Health Statistics. This group recommended a continuing national morbidity survey to provide "needed information on illness in the United States." It identified the potential uses of such survey data as:

- Guide to administrative planning and evaluation of official and voluntary programs in the field of health.
- Evaluation of current morbidity experience in relation to the provision of services, facilities, and personnel for meeting the health needs of the nation or of a community.
- Statistics for medical research.
- Statistics for drug and appliance manufacturers.
- Statistics for public health education programs.

The committee recommended that a continuing survey be undertaken, and to meet the need for data on undiagnosed and nonmanifest disease, that there be physical examinations of a random subsample of the national sample.

The Colombian National Health Survey had its beginnings 10 years later, in 1963, when a group from the two Americas met in New York under the sponsorship of the Pan American Health Organization and the Milbank Memorial Fund. Their charge was to "design an appropriate research approach to the problems of physician needs and of medical education, to discuss the methodology to be used, and to define the appropriate emphasis and parameters of the studies."

This group recommended that one or two Latin American countries be selected in which to undertake, on a pilot basis, a broad series of studies which would include:

- A measurement of the health of the people and health services received in relation to demographic characteristics.
- An inventory of health manpower resources.
- A parallel inventory of supply and utilization of hospitals and other health service facilities.
- An appraisal of educational resources available.
- A study of modifications required in patterns of medical education.
- An assessment of the economic resources available for health services and education.
- Establishment of goals for health achievement and determination of manpower requirements for those goals.

Subsequent inquiry resulted in the selection of Colombia as the pilot country and, to provide the necessary "measurement of the health of the people and health services received", the identification of one of the studies as a national health survey, with both household and clinical examination components.

Thus while the U. S. Health Survey was developed as a more or less freestanding continuing statistical program to meet the needs of a variety of users the Colombian Health Survey was developed as a part of, or an instrument of, a one-time undertaking to establish national health goals, with particular orientation to the planning interests and responsibilities of the Ministry of Health and the schools of medicine.

An original purpose of the United States Health Examination Survey was to identify nonmanifest disease in adults, and thus to complement household survey data. In contrast, plans for the Colombian survey developed on the assumption that in a country with a low level of literacy, and a low level of health services, household interview surveys alone could not provide adequate information on health status. Thus the examination was planned as a rather ambitious health appraisal for all age groups, to include diagnoses based both on clinical and laboratory findings, with some emphasis on acute and parasitic conditions, and on malnutrition.

SURVEY DESIGN

In the United States, the designs of the independent Health Interviews Survey and Health Examination Survey are based on the principle of optimum allocation of resources. Due to differing cost and variance configurations, the designs of the two surveys are quite different. The Health Interview Survey with low unit costs has a large sample of households (about 40,000 annually) in a large sample of places (over 300) throughout the United States. On the other hand, for the Health Examination Survey, for which unit costs were high, a relatively small number of adults were examined, namely 6,500, in a small sample of places (42).

In Colombia, the basic decision to collect information from health interviews and examinations to provide a composite picture of an individual's health status, both as he knew it and as it was known through a clinical examination, implied a small sample of places and persons. Since the unit cost for each person in a health interview survey is low relative to the cost for both interview and clinical examination data (in Colombia, $1.40 as compared to $26.20), it was possible to provide greater reliability for demographic, disability and medical resource utilization data at little extra cost by collecting interview data only for a larger sample of persons than would be both interviewed and examined in each place. The design therefore called for the interviewing of the persons in a sample of about 240 households in each of 40 sample places, either a city, part of a large city or a municipio. A subsample of one out of every 10 persons interviewed was to be examined. Of the 52,480 persons eventually interviewed in 8,960 sample households, 5,260 were selected to receive a clinical examination.

We cannot describe the Colombian sample design in any detail in this paper. A very complete description of the sample has been published in Spanish by the Colombian Health Survey. The design is not unlike that of the Health Examination Survey; it is a complex multistage probability design. Colombia, excluding the territories, vast undeveloped areas comprising 47.3 percent of total land area in Colombia but only 1.3 percent of the population, was first geographically divided into 716 primary sampling units, the units typically being a single municipio, equivalent in many ways to our counties. Because of their large population size, 10 of the 716 PSU's were selected in the sample with certainty. The remaining PSU's were classified into 30 strata according to their elevation, population density, and population size. Next, a sample of PSU's or places were selected from each strata using a controlled technique, the control criteria being geographic proximity and a 4-scale index of health services available. The controlled selection technique differed from other applications known to us in that it controlled for 2 half-samples as well as the total sample. The two half-samples later formed the basis for estimating the reliability of estimates by a pseudo-replication method. One half-sample of 20 sample places, a probability sample of Colombia, was scheduled early in the survey period so that national estimates could be provided if events in 1966 would not permit the completion of all 40 sample places. The probabilities of selection of PSU's were based on current population projections from the 1951 Census. At the risk of some bias, a ratio estimation to 1965 urban and rural population projections from the 1964 Census compensated for deficiencies in the probabilities used in selecting primary sampling units from strata.

By the time we were ready to select the samples of clustered households, advance counts of the number of census schedules or buildings in each 1964 Census Enumeration District had been provided by the National Statistics Department for the 40 sample places. After a conversion of the listed information from buildings to housing units and segments of about 10 housing units, a random systematic selection technique drew a sample of 24 area segments, some urban and some rural, for each sample place. The design called for area clusters containing about 10 households because enumerator assignments in urban areas in the one-day "de facto" 1964 Census averaged 10 households.

Persons to be examined, about 130 in each sample place, were selected from among the persons in sample households, excluding infants. To avoid the necessity of returning to sample households, a method was developed to enable the interviewer to select a sample of persons, stratified by age and sex, at the completion of each household interview. A structured field quality control program, including interviews, gives us a high degree of confidence that the field selection procedures and interviewing were carried out according to instructions.

Once the decision was reached to include clinical examinations in the Colombian National Health Survey, it then remained to devise a statistically sound and operationally feasible survey plan, and then carry it out. An approach discussed but rejected would have the seven medical schools in Colombia interview and examine representative samples drawn from the communities surrounding the schools. Beside the obvious drawback of improper representation of the Colombian population using such an approach, the examination findings would have been subject to large measurement errors arising from the use by the medical schools of varying examination methods, equipment, standards of procedure, and laboratory techniques. The directors of the Colombian survey were quick to realize the need for an operational plan which would reduce measurement errors in examination findings to a minimum. In general, the problem of measurement error can be attacked in two different ways. One approach is to increase the number of measurements, measurers, and measurement devices (the possible sources of measurement error) and by so doing obtain an expected value with a negligible variance. The other approach is to standardize the measurers and measurement devices. In interview surveys both approaches are generally used. Therefore, a standardized questionnaire is used by a number of interviewers trained to conform more or less to standard interviewing procedures.

The U. S. Health Examination Survey attempts to eliminate the measuring device as a source of measurement error by doing the examinations in a mobile health examination center (two were used by Cycle I) consisting of 2 or 3 specially built tractor-drawn trailers, containing X-ray, other examining equipment and laboratory facilities. The mobile center and its calibrated equipment consequently set the standard and provide little or no opportunity for deviation from the standard in any of the examinations made in the mobile center. The cost for this attention to measurement error is obviously high for it sharply reduces the number of places from which persons can be selected for examination and the total number of persons that can be included in an examination survey within a reasonable time period.

In Colombia a mobile examination center was also necessary, since a large portion of the population is remote from any health examination facility. To transport persons selected for the examination survey to the nearest adequate examining facility frequently would require overnight stays and in some instances air transportation. Such inconveniences probably would have resulted in high nonparticipation rates. However, the standardized mobile center approach used in the United States could not be followed in so mountainous a country with such an incomplete road system. Instead, a "standardized examining center" was constructed in each of the 40 sample places. Existing health centers, clinics, or hospitals in or near the sample places provided the walls. Most of the examining equipment and the examination technicians were moved from place to place by whatever transportation means was required. In some cases this meant a combination of air, land and water transportation. Five clinical teams and sets of standardized examining equipment were in operation throughout most of the 8-month survey period. In amazing feats of logistics each clinical team packed-up from one examining location, set up in the next location about 4 days later, and after typically two weeks of examinations packed-up again.

DATA COLLECTION

Interviewers in the United States National Health Survey, all women, are part-time employees of the Bureau of the Census. No prior education or work-experience in the health fields is required of the interviewers. Since the housewife is usually the respondent to a household interview in the United States, it is believed that female interviewers obtain a higher degree of cooperation and more complete reporting. Contrastingly, interviewers in the Colombian Health Survey were medical students, most of them male. The traditional labor pool in the United States--educated housewives seeking part-time employment--does not exist in Colombia. Medical students were used in Colombia not because of any conviction that they would be superior interviewers, but because they represented the best available manpower supply. Since medical students could be away from classes for only 15 days at most, a new set of 6 interviewers were trained and employed in each of the 40 sample places. Another important, nonstatistical reason for using medical students is closely tied to the origin and objectives of the Colombian Health Manpower and Medical Education Study. It was believed that participation of medical students as interviewers, and residents in internal medicine and pediatrics as clinical examiners, would offer a unique opportunity for medical students and young physicians to actually get out into the country and to see what people were like in their homes and what their health problems were in relation to the places in which they lived. The Colombian medical schools have since indicated that the kinds of perception gained by medical students in these field operations are already having substantial impact on the teaching in their institutions. As yet there is not sufficient evidence to evaluate the performance of medical students as interviewers. The high response rate (97 percent) indicates that there was little difficulty in gaining cooperation of sample households.

Turning to the questionnaire of the Colombian Survey, again we find that conditions peculiar to Colombia and perhaps other developing countries were as important in determining its composition as were the requirements for information. At the outset George Kearns, the consultant on interview survey methods, recommended against any wholesale adaptations of the Health Interview Survey questionnaire. The latter is an unusually detailed and complex questionnaire that requires weeks of instruction and practice to bring new interviewers up to the level of satisfactory performance. In Colombia both the interviewers and their supervisors were totally without prior interviewing experience. It was extremely doubtful that their level of performance could be brought up to that of their U. S. counterparts in a short time. So, while the Colombians retained many of the basic features of the Health Interview Survey questionnaire, they discarded one of its principal and most difficult elements--that series of questions which seeks to obtain detailed characteristics of each reported morbidity condition, subsequently coded to the 4-digit ICDA code. Instead, a modest effort was made to measure prevalence of only nine of the more common chronic conditions recognizable even by an untutored or illiterate layman.

The sections of the United States questionnaire retained with only minor modifications were those that obtained information on demographic and socio-economic characteristics; incidence of acute illness and injuries, with the standard two-week recall; restricted activity due to illness; "doctor" visits (and within that, visits to, or service by "quasi" health practitioners such as midwives, pharmacists, teguas, and so forth); amount and source of payment of the "doctor" bill; and dental visits. There was also a section on hospitalization which included questions on cause, length of stay, and amount and source of payment. The recall period for hospitalization was not the traditional "last 12 months"--a difficult concept at best, even here in the United States. Instead it began with a fixed almost universally celebrated and remembered date in Latin America, depending on the month of interview either New Year's Day or Holy (Easter) Week. Because of the varying interview months for the 40 sample places, an adjustment factor was used to convert the data from the recall periods of from 9 to 15 months into annual estimates. As far as we know this caused little trouble in execution and hopefully reduced the response error experienced when using the "last 12 months" recall period.

As to the mechanics of the form, we like to think of it as a model of simplicity. We know from reports of the observers that it presented few difficulties even to the marginal interviewers. It was highly structured, easy to read, with uncomplicated skip patterns, and with sufficient space for recording answers. The format of the main part of the questionnaire was particularly efficient. It was a booklet with provision for a seven person family -- with facing pages for each person. Multiple events such as hospitalizations were handled by using additional "person" pages for each event. The left third of the page contained instructions to the interviewer, the central third contained the actual questions, while the right third was the answer section including a shaded code area for those categories which could not be pre-coded. This eliminated the necessity for separate transcription and simplified the punching operation. Undoubtedly some improvement could have been made in the questionnaire. However, despite the relative inexperience of the project staff, it was one that was carefully phrased, logically put together, and one which worked well both in the field and at the processing stage.

The clinical examination followed the U. S. protocols, except for the emphasis on acute and parasitic diseases and nutritional status. Procedures were carefully standardized, as in the United States survey. Special techniques were developed to assure accurate laboratory work. Because of the short supply of technicians, the Colombian National Institute of Health provided a 3-month training program for technicians for the clinical teams, and made the laboratory determinations for a number of the tests.

DATA PROCESSING

Up to this point our comparisons have probably been prejudiced towards those areas which cast the Colombian survey in a favorable light. As should be expected for an undertaking of such dimensions, some problems were encountered for which no satisfactory solution was found, and some mistakes were probably made in planning and executing the survey. Naturally, we tend to forget most of these. However, what was probably our greatest failure as senior consultants continues to haunt us today. To date only a fraction of the information contained in the Colombian National Health Survey has been tabulated, particularly with regard to the examination findings. Although two excellent reports on findings from the interview survey have been published, as have several methodological reports, progress has been slow. Our failure was one of not giving more emphasis to data processing requirements at an early stage, although we included a data processing consultant on the staff of U. S. consultants from the beginning.

The primary problem encountered in data processing was not one of sheer volume of data, but rather a lack of the advanced programming skills and tabulation control procedures necessary for wringing out correct tabulations from the raw data of a complex sample survey. This is not a problem peculiar to Colombia, but one which is widespread among developing nations, not to mention new programs in the United States. The U. S. National Center for Health Statistics has developed the needed programming skills and control procedures over a period of years of training, practical experience and a close working relationship with the Bureau of the Census. We allowed ourselves to believe that these skills had already been developed in the National Statistics Department (DANE), the initial data processing agent for the Colombian National Health Survey. We believe it is correct to say that as yet neither DANE nor the present contractor have developed to that stage. In the meantime programs of the U. S. National Health Survey have been modified and used in Colombia to produce marginal tables.

We feel that the problems we are now encountering could have been forestalled if: a person with basic programming skills had been added as a full-time member of the Colombian staff and trained for 3-months at the U. S. National Center for Health Statistics; and the entire Colombian staff had been better indoctrinated in the need for and methods of tabulation control procedures.

SUMMARY

At the October 1967 Round Table meeting of the Milbank Foundation, convened to review the methodology of the Colombian National Health Survey and its implications for the social sciences and health planning in Colombia, one of the participants asked "Scientifically, what is new in this report?" He was referring to the findings, in his belief that they simply confirmed findings already well known. However, the question could also be directed at the methodology, and in a sense was because the methodology generated the findings. The answer to the question "what is new?" would then be the demonstration that national sample surveys can be carried out in a Latin American country despite very formidable obstacles. Among the many deep satisfactions we have gained from our Colombian venture is the knowledge that science has been advanced in Colombia and to some extent outside Colombia. The Colombians have been exposed to probability surveys, the methodology is now well-documented in Spanish for others to profit from, and a nucleus of experienced survey technicians has been formed. The methodology of the Colombian survey has been extended to a dental survey, now in the field in Venezuela, and has influenced the planning of health surveys in

Argentina and Brazil. Finally it has been shown that a program developed for national health planning purposes can also provide the basis for international comparisons in morbidity and health services.

REFERENCES

(1) Colombian National Health Survey: Planning, Methods, and Operation. Reprinted from the Milbank Memorial Fund Quarterly, April 1968, Vol XLVI, No. 2, Part 2, New York, N.Y.

(2) Gomez, Luis C. and Losee, Garrie J.: La Muestra de la Investigacion Nacional de Morbilidad. Bogota, Minasalud, 1968.

(3) Health Manpower and Medical Education in Latin America. Report of a Round Table Conference. Reprinted from the Milbank Memorial Fund Quarterly, January, 1964, Vol XLII, No. 1, New York, N.Y.

(4) U. S. National Health Survey. Origin and Programs of the U. S. National Health Survey. Public Health Service Publication No. 584-A1, Washington, D C., 1958.

(5) U. S. National Health Survey. Plan and Initial Program of the Health Examination Survey. Public Health Service Publication No. 584-A4, Washington, D. C., 1962.

(6) U. S. National Health Survey. The Statistical Design of the Health Household-Interview Survey. Public Health Service Publication No. 584-A2, Washington, D. C., 1958.

VIII

VOCATIONAL EDUCATION STATISTICS

Chairman, MORRIS B. ULLMAN, U. S. Office of Education

THE NEEDS FOR VOCATIONAL EDUCATION DATA

by Morris Ullman, Bernard Michael and Marc Matland [1/]

Introduction

In this period of rapid change, society is constantly making new demands on our always limited public finances. We do not have the resources, however, to satisfy all these needs and although we live in the most promising period in recorded history discontent is growing. To resolve the problems facing society with perpetually limited resources makes rational planning and increased efficiency imperative.

In education, as in other areas of public concern, and in vocational education in particular there has been a demand for improvement and exploration. In response, we are probably developing and implementing more innovation in occupational education than at any time in the past.

But the multiplicity of experiments requires careful evaluation if we are to separate the meaningful from the inconsequential and succeed in improving vocational education and making it relevant to the needs of the individual and the community today.

To even attempt to implement the broad mandate for a meaningful vocational educational system with available resources demands effective program planning, program management, and program evaluation. Identifying information needs for planning more efficient and effective occupational education programs is, therefore, one of the first orders of business.

Defining Vocational Education

"In a broad sense, all education contributes to vocational competency, but vocational education refers to that part of a student's instruction intended specifically to fit the student for work."[2/] For the purposes of this paper, the term "vocational education" refers to all formal instruction for both youth and adults, at the high school, post high school, and out-of-school levels, which prepares individuals for initial entrance into and advancement within an occupation or group of related occupations. The function of vocational education, then is twofold. It must provide the people it serves with an education and it must train skilled workers for the labor force.

Training in vocational education falls into seven major occupational program areas: agriculture, distributive, health, home economics, office, technician, and trade and industrial. Within these seven broad categories, there are more than 130 areas of instruction designed to meet the needs of both the individual and the labor market. In agriculture, for example, there are approximately eight instructional areas, while in health occupations there are somewhere in the neighborhood of 18 areas. (Others are: technician, 26; trades and industry, 40; distributive, 20; home economics, 8; and office, 11.) The vocational training here defined is generally conducted in public institutions, administered and financed largely by the States and local school districts, and encouraged and supported by the Federal Government.[3/]

Briefly, our public vocational education systems involved, in 1967, seven million students--half of them at the high school level, 504,000 in postsecondary full-time programs, and almost 3,000,000 more in adult, part-time programs. In the system, these students overlap with regular high school enrollment, with junior and senior college enrollments and comprise a significant number of enrollees in adult education activities.

Reports from the States indicate that the expenditures in 1967 for the vocational education programs was a shade over a billion dollars, was slightly over one-quarter (26.5%) being provided by the Federal Government.

The growing impact of the Vocational Education Act of 1963 can be illustrated by the fact that since 1963 the Federal expenditure has increased fivefold and the State and local expenditure has increased threefold. About $200 million of the 1967 expenditure was for the construction of new facilities.

The expected passage of legislation now pending will increase the level of Federal spending on these programs. On the basis of past behavior, State and local expenditure will also increase by at least similar amounts.

Relation to Other Occupational Training Programs

In putting these limits on our discussion of vocational education we do not intend to ignore other occupational training programs which make important contributions to our labor supply: courses in proprietary schools, projects of other Federal agencies, including the antipoverty programs and professional and technical training in the Armed Forces; the training conducted by business and industry; and other sources. In numbers, however, the vocational education programs make the largest contribution. Other Federally sponsored programs, such as those under the Manpower Development and Training Act, the Job Corps, the Office of Economic Opportunity, the military's Operation Mainstream and New Careers, have important impact, but these programs tend to concentrate on more difficult problem areas and involve smaller numbers than do the regular vocational education programs discussed here.

Since the present studies of labor supply and demand are largely institutional and somewhat general and since there is considerable geographic and occupational mobility, it is difficult to assess the contribution of vocational education to labor supply and to economic growth. Research aimed at such an assessment for specific population groups is now underway. A question on vocational training for the 5% sample of the 1970 Census is planned and, if plans persist, will provide information on vocational training in the background of the working population. The information furnished will permit further analysis of the long-term economic and social impact of vocational education programs.

Important revelations on the past relationship between education and occupation comes from the reports of the decennial Census of Population. In addition to the significant tabulations on educational attainment by occupation from the 1960 Census, a special supplement to the Current Population Survey in April 1963 provided information on the occupational training of Adult Workers.[4/] Similar surveys in connection with recent Federal manpower and training programs obtained information for 1966 and 1967 and these data will soon be forthcoming.

Relation to the Total Educational System

Vocational education, as we have defined it, is an integral part of our public educational system and subject to the same pressures as affect the system as a whole. The increase of emphasis on vocational education is part of the demand for more effective education in general. This emphasis is reflected in the fourfold increase over a ten-year period in all Federal expenditures for education and training, in a sevenfold increase in the programs administered by the Office of Education, and in a sixfold increase in Federal funds for vocational education.

As the Advisory Council on Vocational Education points out, it is no longer possible to compartmentalize education into general, academic, and vocational components. Education is a crucial element in preparation for a successful working career at any level. With rising average educational attainment and the changing technology which requires it, the less educated find that the opportunities for employment for them are becoming more scarce. On the other hand, employability skills are equally essential to the education of the individual. If education is preparation for life, and if practically everyone's life and opportunities for self-expression and self-fulfillment include work, then only the successfully employable are successfully educated.

As a consequence of these attitudes and the evaluations they require, it must be recognized that vocational education statistics without data on other education, will never be completely adequate. Vocational programs and students must be identifiable within the total educational framework. Definitions must apply to the entire system to properly measure any segment of our educational effort. Much of our present data suffer from this lack of common identification and comparability between various segments of our education community.

Coordination of Vocational Education with the Labor Market

With the enactment of the Vocational Education Act of 1963 and the recent amendments, the Congress of the United States has left no doubt about its concern for establishing a flexible vocational education system with services available to all Americans of all ages, of all backgrounds, and of all levels of ability, and in all communities. Special emphasis is also placed on providing high quality vocational training, or re-training, which is realistic in terms of the opportunities for employment. Federal funding in the new bills now in conference calls for approximately double the 1968 authorization in 1969 and substantial increases thereafter.

In planning vocational offerings, administrators at Federal, State, and local levels must consider the needs and desires of the individual, the community in which he lives, and the private and public agencies which might employ him. The planners must know the characteristics of the target populations to be served and of the local areas where they live and work. The planners must also be able to determine how local needs fit in with the requirements of the State, the region, and the Nation as a whole.

We need labor market information which will permit planning of vocational programs to be responsive to the needs of industry and to be useful in preparing students for viable occupations and careers. Information needed includes a comprehensive description of the labor market structure in the local area and how it is changing. Data are needed to indicate the magnitude of the demand for workers in each occupational group and the nature of the demand--how much is due to turnover, or to growth, what is the probable impact on the demand by sex and age groups. We must also know the extent to which the occupational groups are diminishing or expanding, and the major factors influencing the changes.

Educational planners, however, must consider the requirements, desires, interests, and abilities of the individual, as well as the needs of the economic system. Basic manpower data of the type described are only the beginning. Education and training programs must be provided for various levels--those designed to prepare individuals to enter specific jobs, and those aimed at providing a basis for the eventual handling of more complex tasks, as well as provision for re-training, and for up-grading. More information is needed on the education and skill requirements of related job families at various levels, entry and beyond.

Moreover, there is urgent need to tie in occupational projections with information about job content and projected changes in job content. Educational planners must see far in advance where major changes are likely to be made because of the lead time involved in developing new curricula, in organizing programs and in recruiting and counseling students. In some cases, planners will have to estimate magnitudes and commit funds for the design of buildings and facilities, and for the purchase of machinery and equipment that will affect educational and training offerings far into the future. Plans for updating teachers or for hiring new teachers also have to be made far in advance. And for public education, allowances also must be made for the legislative requirements for funding and implementing any decision reached.

The Vocational Education Act of 1963 recognized this situation and provided for joint planning efforts by the Departments of Labor and of Health, Education, and Welfare. Reflecting this, local offices of the Bureau of Employment Security were called upon to furnish projections of needs so that educational agencies can plan activities to meet these needs. Data on both needs and training plans are to be summarized in an annual "Projected Activities Report." The total effort needed to provide such data is tremendous and only the first efforts are now visible. The Labor Department has prepared guidelines for making such projections. The Office of Education has prepared a taxonomy of training programs. A manual is now in preparation which will provide a bridge between occupations listed in the Directory of Occupational Titles (DOT) and the educational programs defined in the OE Handbook, Standard Terminology for Instruction in Local and State School Systems.

Such steps, of course, are only beginnings. In order to determine what is to be taught, the best way to teach, and how to evaluate progress toward reaching objectives, much more data are needed on relationship of training to performance. Data on the characteristics of education (organizations, curricula, facilities, teachers, instructional methods, services to students) must be related to employment (placement and job performance) of the graduates from vocational education programs. To do this adequately, we need to know something about the personal and educational characteristics of these individuals--their abilities, aptitudes, educational levels, occupational training received, socio-economic origins, and performance in school. We need such information not only for full-time preparatory students, but also for adult trainees. We need to know much more about those who drop out of vocational programs and also about those who might have desired, but were unable to obtain, vocational education.

We know very little about the characteristics of the "disadvantaged" or "special needs" groups served by the vocational education programs or even precisely how persons included under the umbrella term "disadvantaged" should be identified so as to obtain meaningful information about them. Because of the variety of groups included, we must establish meaningful categories for study. For example, what are the similarities and differences between central city Negroes, Spanish-speaking Americans, and American Indians? What type of data would enable us to marshal our resources to provide them with the most meaningful training? What are the major obstacles to successfully bringing these groups into the mainstream of society? What use has already been made of Federal and other funds to help them? In short, how should we measure effectiveness and efficiency of programs for the disadvantaged?

More information is needed also about teachers, both short- and long-range supply and demand. The effectiveness of different types of teacher preparation needs evaluation in terms of both teacher and student performance.

We also need more information which cuts across programs. For example, to what extent do persons who receive training under the Manpower Development and Training Act or in Adult

Basic Education move into regular adult vocational classes or into other channels to continue their training? Because of various joint arrangements involving two or more programs, the little data we do have sometimes lose meaning, because we know there is duplication, but we do not know how much.

The Evaluation of Vocational Education Programs

Effective management of vocational programs whether at the Federal, State or local level, also requires an information system which will facilitate progress evaluation. If evaluation is viewed as the basis for improving administration of programs, the measurement phase on which judgments are based becomes especially important. In this sense, measurements include not only an established information system, but also special studies to supplement the basic system. Considerable effort is needed to strengthen concepts and to improve collection methods in order that both general and specific facts gathered may be more meaningful. If an objective is to study the efficiency of programs, more has to be done to determine just what should be included as measures of costs and benefits and how much weight to give them. For example, if funds were allocated on the basis of economic efficiency alone, all vocational education dollars available might be invested in post-secondary programs which, by certain criteria, yield the greatest marginal return per extra dollar invested. If this were done, funds for secondary vocational programs would be reduced. Such a decision would ignore completely the fact that the rate of teen-age employment is higher for those without vocational education and that the costs of aggravating this situation could be reflected both economically and socially.

The administrators and the school boards, faced with the problem of allocating limited resources, have difficult choices to make. Many factors enter into their decisions which become the objectives for educational programs. It may be observed that under the conditions that usually prevail, cost benefit comparisons are most useful for making choices between alternative programs serving the same group and must not be used for inter-group comparisons.

Congress has specified that considerable emphasis be placed on programs for the disadvantaged, particularly for young people who do not succeed in regular vocational programs. Such special-needs programs often require remedial services such as basic literacy instruction, counseling in depth, medical services, job development, and similar services. Meeting these needs may also require the development of special curricula and methods. To evaluate this type of program our knowledge of what benefits to measure and how to measure them needs further analysis.

More needs to be known about what impact Federal support has on vocational education programs. Planning activities and policy determination at all levels of government are often based on evaluations of what <u>they</u> are getting for <u>their</u> money. The examination of the effectiveness and efficiency of the choices available to those responsible for allocating funds does not always consider issues broader than their immediate concern. Local issues may have priority over matters of national concern.

We have already noted that Federal funds in this area are more than matched by State and local funds. What would happen if Federal support for particular programs was reduced or withdrawn? Which programs would continue to flourish and which would begin to wither? Or to raise these questions from a different viewpoint--what are the key leverage points for change? Where would additional support make the greatest difference--teacher preparation or improvement in State planning? Would general aid to education be less or more effective than categorical aid to specific programs?

To effectively evaluate progress of vocational education more data are needed on outcomes of the programs. We need more follow-up information on individuals--not just on initial placement, but on job satisfaction, further education, advancement, mobility, employer satisfaction. This information should be related to the backgrounds and characteristics of the individuals.

The small amount of outcome research already available to us has been focused largely on what has happened to those persons completing vocational programs. Few of these studies have been designed to compare what happens to similar groups of students, as for example, to graduates of high school general or academic programs, or to those whose occupational training is entirely on-the-job. Techniques need to be developed which will allow us to evaluate various approaches to training. How do we measure public school vocational education in relation to other training programs--MDTA, Job Corps, and other OEO programs, welfare training programs, OJT? What are the relative costs of each type of program and which offer the greatest benefits?

We also need measurements to help us assess the relative effectiveness of vocational and other programs in attracting and holding as well as placing students. We need to know more about vocational education as an alternative approach to learning as well as for skill development.

Finally, we must mention the need to evaluate the impact of vocational education on the broader social problems of the day--the urban problem, rural poverty, specific skill shortages, discrimination practices, and others. For each such problem area, the role of vocational education needs to be identified.

Data Available

Current national data on vocational education are produced (1) from the annual reports of the States to the U.S. Office of Education, (2) from broader population studies, either general or educational, such as the Census Bureau's Current Population Survey, AIR's (American Institute for Research) Project Talent, and the Equal Education Opportunities Survey and (3) from various research studies usually limited in scope and directed to a specific problem in vocational education. Data on less than a national basis are also available in State reports and in other special studies. Since our emphasis is on national data, the State studies are not covered here, but they should not be ignored as sources of important information.

The principal national data sources are the reports from the States to the U.S. Office of Education, summarized in the annual report, Vocational and Technical Education.

These annual reports are submitted by the States each fall and consist of summaries, aggregated to the State level, of the previous school year's finances, enrollment, programs, teacher training, and other activity. States are requested to summarize in these reports the combined figures for all federally reimbursed programs and for all non-reimbursed public programs meeting the same standards. The 1963 Vocational Education Act extended coverage of this reporting program to all types of occupational training and, therefore, made possible reporting of data on all occupational training in the public institutions of the State. One of the important tasks for the immediate future is the assessment of the quality of these data.

In addition to regular summaries of finances and enrollment by program, two reports recently added to the annual series might be mentioned. About three years ago a form was added asking for the employment status in October of those who had completed full-time programs during the previous school year ending in June. The results are available only in State summary form and for the seven standard vocational programs. For all completions and for the completions by programs, information is available on such items as the extent to which those persons completing programs continue full-time education, the extent to which they are available for employment, and the initial employment and unemployment rates.

Also about three years ago a form was added on the States' projected activities in vocational education. Previously some of the planning had been reported in narrative form. The present form is developing into an instrument for planning vocational programs. Space is provided for projections of labor demand in pertinent occupational fields to be furnished by the State Employment Security Office; enrollment projections; and related information. Information obtained on these forms has already been used for legislative and other reporting. However, it must be recognized that it will be some time before the information systems needed to process these types of data will be adequate for more refined use. Both the Department of Labor and the Office of Education are preparing materials and working with the States to improve their ability to present this type of information. The Advisory Council has recommended that this report on projected activities replace the present State plan, that it be prepared with great care about once every five years, and that it be updated by the States each year.

The above reports give us no information on individual schools or on areas smaller than the States. Where analyses require such data, as in cost benefit analyses, research workers have had to engage in their own data collection on a special study basis, as is the case in the Pennsylvania State University studies reported elsewhere at these meetings.

Also missing from the State reports to the U.S. Office of Education is any information about characteristics of individual students, teachers, or programs. This need was noted in very specific terms in the 1968 report of the Advisory Council. To overcome this lack, in part, a project was undertaken by the Center for Studies in Vocational and Technical Education at the University of Wisconsin. With the cooperation of the State offices this project has already produced a Directory of Vocational Education Programs, 1966, which will be used as a basis for surveying graduates and drop-outs in secondary and postsecondary institutions.

Some data on vocational students can be extracted from broader educational studies, specifically Project Talent and the Equal Opportunities

Survey. The major problem in using such data is the definition and identification of the vocational education student. Most secondary vocational programs are given in comprehensive high schools or other schools along with other programs and the students' responses as to their principal programs may differ from those of the school. In the Equal Educational Opportunity Survey, 97,000 twelfth grade students were included in the sample. There were three ways of identifying vocational students: (1) self-identification as to their programs, (2) coursework taken in certain fields in half-year intervals, and (3) identification of the schools as vocational schools by the principals. Only about 40 percent of the responses were fully consistent. 5/ Work is still underway with this body of data to try to clarify a consistent method for identification of students in this area.

Data on teachers will be available from another research project being conducted by the Bureau of Social Science Research. Other studies are available on specific areas of vocational education such as the study of graduates of trades and industrial occupation programs by the American Institutes for Research. Products of such research are reported through the U.S. Office of Education's Education Resources Information Center (ERIC). In this network, the clearing house for vocational and technical education is the Center at Ohio State University, which publishes quarterly "Abstracts of Research and Related Materials in Vocational and Technical Education" (ARM).

Plans for National Data

Present plans of the U.S. Office of Education for national statistics on vocational education are focused on improvement of current reporting; supplementing existing reports with other needed series such as those on characteristics of individual students, teachers, and programs; sponsoring special studies for information for program planning and evaluation; and continuing the sponsorship of research which will not only provide new information but will also help to sharpen concepts and definitions for measurement.

Current reporting will be affected by legislation now under consideration. The revised legislation should permit simplification of financial reporting. A thorough revision of the annual national education reports is expected this coming year. At that time it is intended to explore the possibility of collecting data by major cities and other geographic areas.

Much work has gone into the development of a taxonomy of educational programs. OE Handbook VI, "Standard Terminology for Instruction in State and Local School Systems," names and defines the vocational education programs and has been used experimentally in reporting for the last two years. The comparability manual now in preparation, which will relate the OE Instructional Codes to the Dictionary of Occupational Titles, should provide an essential tool, not only for statistical reporting but for program planning, educational counseling, and many related activities at all levels of operation. 6/

Annual reporting on vocational education by the States should also benefit from another manual, now in preparation in the U.S. Office of Education which will give more specific instructions for filling out the required forms. Concepts will be related more closely to those currently in use in the general educational system as well as to labor market analysis and to State practices. The use of this manual should improve the comparability of reports from State to State.

To supplement the State reports, the U.S. Office of Education hopes to launch this winter, a sample survey which will cover the characteristics of students, of teachers, and of programs. The exact plans for this survey are dependent upon the availability of resources which have not yet been determined, but it is anticipated that a substantial start will be made during this fiscal year to inaugurate a regular series.

A series of evaluation studies is also planned but again, the extent of these studies will depend upon resources made available in the pending appropriations as well as upon competing claims for these resources. Now underway are such studies as the follow-up and characteristics of graduates and drop-outs from various types of vocational programs. Also underway are a number of cost benefit analyses to develop information on relative values of alternative programs, staffing patterns, and sizes of school districts.

We have mentioned more data needs than can possibly be produced in a single program or in a short period of time. Neither will we be able to satisfy all the Federal, State, and local needs at the same time. It is important, however, that we develop some perspective on our objectives so that our resources can be directed most efficiently. If this paper indicates some of the problems toward which our statistical resources might be directed, or if it stimulates more research in these areas, it will have achieved its purpose.

1/ The authors, all in the U.S. Office of Education, are in the National Center for Educational Statistics; the Bureau of Adult, Vocational and Library Programs; and the Bureau of Research, respectively. The paper reflects their personal experiences.

2/ Education for a Changing World of Work, Report of the Panel of Consultants on Vocational Education, U.S. Department of Health, Education, and Welfare, Office of Education, U.S. Government Printing Office, Washington, D.C., 1963, p. 5.

3/ The description of this local-State-Federal financing system and summary State data appear in U.S. Office of Education, Vocational and Technical Education Annual Report, Fiscal Year 1966, Document OE-80008-66, Washington, D.C., 1968. The system is reviewed and evaluated in the 1968 general Report of the Advisory Council on Vocational Education which is published in the 90th Congress Committee Print, "Notes and Working Papers Concerning the Administration of Programs Authorized Under Vocational Education Act of 1963, Public Law 88-210, as Amended," prepared for the Subcommittee on Education of the Committee on Labor and Public Welfare, United States Senate, March 1968.

4/ U.S. Department of Labor, Formal Occupational Training of Adult Workers, Manpower/Automation Research Monograph No. 2, Washington, D.C., Dec. 1964. See also U.S. Bureau of the Budget, Current Occupational and Past Training of Adult Workers, Statistical Evaluation Report No. 7, Washington, D.C., March 1968.

5/ Leslie J. Silverman and Katherine K. Wallman, "The Identification of Vocational Education Students in the Equal Educational Opportunity Survey Data" (multilithed), U.S. Office of Education, Washington, D.C., 1968.

6/ A specimen page of the comparability table is included as Appendix F in Annual Report for Fiscal Year 1966, Vocational and Technical Education, U.S. Office of Education, Washington, D.C., 1968.

THEORETICAL AND EMPIRICAL PROBLEMS IN THE ANALYSIS OF THE ECONOMIC COSTS OF VOCATIONAL EDUCATION

Ernst W. Stromsdorfer, Teh-wei Hu, and Maw Lin Lee,*
The Pennsylvania State University

A. Introduction.

Within recent years there has been increasing interest in analyzing the economic costs and benefits of vocational education.[1] Yet this entire discussion has largely been carried on either in an informational vacuum or in contexts where gross analytic errors have been made. For instance, there is a general lack of awareness that, in order to make efficiency judgments, the extra or marginal costs to vocational education must be related to the extra or marginal costs of vocational education. Often, programs of a given cost are asserted by their detractors to be "too expensive" while the program advocates see these same costs as evidence of "high quality." Such assertions are in error. Other analyses utilizing benefits either do not relate costs to benefits or else measure gross instead of net benefits of vocational education. No valid policy statements can be made from such analyses, yet the policy statements are made.

Compounding these problems is the fact that rudimentary data needed to make decisions at a relatively gross level generally do not exist. This is due in part because it is not reasonable to expect a school district to expend resources to collect data it either sees no use for or has not been educated to use. There is also the problem that the data needs of decision makers at different levels in the educational system do not coincide since the problems which face different levels do not coincide. It has not always been accepted or perceived that the financial and educational system data being collected ought to coincide with the needs of educational decision making. Finally, the acceptance of rational economic analysis of the educational process if relatively recent.

This paper describes the cost data availability, needs and problems encountered in an attempt by the authors to perform an economic evaluation of vocational-technical education.[2] Cost analysis is used to illustrate data problems. Next, an evaluation is presented of national cost data on vocational education from the standpoint of the needs of economic analysis. This data is used to illustrate the econometric problems involved in attempting to employ it in cost benefit analysis.

B. Problems in Cost Analysis.

All costs must be considered as opportunity costs. Any complete analysis of costs should measure both social and private costs, and, in some cases, the costs of governmental units.

Under social costs, the following should be considered:

1) Current costs, which include such factors as teachers' salaries, heat, light, and other variable costs;
2) Capital costs of sites, buildings, and equipment;
3) Cost correction factors such as sales tax and property tax correction factors;
4) Costs from nonschool system support;
5) Earnings foregone while students are undergoing education;
6) Incidental costs to students associated with school attendance;
7) Job search costs; and
8) On-the-job training costs.

Under private costs the following should be considered:

1) Earnings foregone while the student is undergoing education;
2) Incidental costs associated with school attendance;
3) Job search costs; and,
4) On-the-job training costs.

None of these costs are conceptually different. They are listed separately because each has measurement problems peculiar to it.

Total, average, and marginal costs should be measured. These costs should be related to specified production functions. The production functions should incorporate those input variables which affect their determination and structure, such as class size, or number and quality of teachers. Cost-benefit analysis is concerned with the making of decisions which allocate resources efficiently, so that the main concern of this analysis is with the determination of marginal, or extra costs of producing an additional unit of output. Problems of cost determination occur with respect to measurement of total and average costs of a given output or set of outputs when joint costs occur.

The Joint Cost Problem. Joint costs occur within two contexts. First, the problem exists at a given point in time when a specific educational input or facility produces two or more distinct educational outputs. Second, the problem occurs over time, when a facility is consumed during the investment or training process by successive cohorts of students representing either the same or a different type of output.

Fortunately, the occurrence of joint costs does not affect the determination of marginal costs. And, since efficient investment decisions between two or more alternative projects are made on the basis of marginal costs, the presence of joint costs presents no basic problem to cost-benefit analysis.

In actual practice, however, costs which are joint are frequently allocated among different programs. Not only is such allocation always arbitrary in nature, but it is unnecessary. When joint costs occur, the total cost of the set of programs or outputs combined can be measured. Then, their combined benefits must equal their combined total costs. But total average costs to each of the programs simply cannot be measured accurately in any economic sense.[3]

Consider the following: Both vocational and nonvocational training occurs in a comprehensive senior high school. In this school certain costs are directly attributable to a given program in vocational education, such as the extra costs of electricity to run the power tools of the machine shop or the extra wiring installed in the shop room. However, the building itself needs a given electrical system to feed electricity to all the various classrooms and shops. This cost outlay serves both the vocational and nonvocational students. Given that a decision has been made to install a machine shop in that school, no part of the common cost of constructing the basic school building should be included as a cost offset to the benefits flowing from the machine shop. The correct allocation of these common costs to the machine shop operation, and by extension, the costs of training students as machinists, is simply, zero. This is so because, within the limits of the feasible range of output in the school, the use of the common facilities by the students taking machinist training does not reduce the ability of the other students in the school to use the same common facilities. Thus, joint inputs are similar to a public good which anyone can consume as much as he wishes without reducing the consumption of that good by others.

In this regard, then, Handbook II, Financial Accounting for Local and State School Systems, is in error.[4] It advocates the proration of such inputs as administration, attendance and health services, transportation services, operation of plant, maintenance of plant, fixed charges, food services and student body activities and community services, all of which are joint in nature or have joint components.

Current Costs. Some current costs will be specific and some will be joint. Given a comprehensive school which produces more than one type of product or provides different types of specialized training, typical joint costs could involve the items listed above. Even if, as with the school lunch programs, students were charged a fee which reflected the cost of providing lunch to each of them, differences in marginal cost between different students would not necessarily be affected, if, as is often the case, a flat fee would be charged to each student. Of course, one would attribute as a cost of education only those costs involved in food preparation and serving which would be over and above what the student would normally incur were he not in school.

Controversy exists over whether or not such in-school programs as attendance and health services and community services represent aspects of the educational process. In some respects these programs are similar to other public health and social services and an argument could be made for including such expenditures in their respective community-wide programs. However, there are interaction effects between the state of one's health, nutrition, and quality of life and the educational and learning processes. So, total exclusion of such expenditures in an accounting of the costs of education may not be warranted, since these programs do facilitate the educational process. A case could be made, however, to attribute the increased effectiveness of the educational process brought about by such things as health expenditures as a benefit accruing from the health program. Our judgment would be to exclude these expenditures whenever possible and recognize that in their presence, the benefits accruing to the educational process per se are overestimated.

Specific costs could involve such matters as the cost of the shop or classroom teacher, the cost of supplies and books associated with a given educational curriculum, or maintenance or janitorial services associated with each curriculum. Clearly, the cost accounting problems associated with maintaining a separation of the joint and specific cost aspects of a given input are severe. Thus, to some extent, a counsel of perfection is being recommended.

Capital Costs. Social (and private) capital costs are fundamentally no different in nature than social (and private) current costs, and what follows should not be construed as suggesting so. To the extent that costs are categorized, this is done because each type presents different measurement problems. Capital costs can be broken down into four different elements:

a) Site acquisition costs;
b) Capital improvements to the site;
c) Physical plant and building costs; and,
d) Equipment costs.

Serious measurement problems stem from several physical and institutional factors. Two of the most important factors are: (1) The physical plant of the school usually has an economic life longer than the period of training for any given educational cohort; (2) the services of this capital stock are not easily valued in market terms.

Four possible treatments for valuing this capital exist. First, one can argue that once the capital stock exists, especially the physical plant and buildings, it becomes specific to the educational process and thus has no alternative use in the short run. This is a tenuous assumption, though, for it is easy to discover alternative uses for such capital stock. Thus, the value of the educational physical plant is not zero, but since it is not a perfect substitute for competing uses, the value of the competing uses, such as the rent of a hospital, does not reflect the exact opportunity cost of using the physical plant for educational purposes.

Second, historical costs of building construction and site acquisition can be used, but these historical costs are irrelevant since they have no necessary bearing on the present opportunity costs involved in using any capital stock.

Third, the use of replacement costs is a possible measure of capital costs. However, it is obvious that often it would cost more to exactly replace a building than the current economic value of the building. The use of replacement costs would over-value the capital resource, given a rising price level and assuming no compensating technological change in construction technique.

Fourth, an estimate of current assessed valuation could be used to arrive at a measure of the capital costs. However, the valuation standard used becomes critical. In actual practice, the valuation standard amounts to a combination of historical costs adjusted by a price index of replacement cost so that this measure is no better than the replacement cost measure. Unfortunately, this is essentially the practice followed in two of the three cities in this study which do report assessed values of their buildings and physical plants for purposes of fire insurance.

In short, it is not obvious what price resulting among these four choices should be attached to the capital inputs to get a measure of the opportunity costs.

The Capital Recovery Factor. Once the economic value of the capital in use has been measured, a problem remains with the measurement of the rate at which the given capital stock is consumed over the course of the investment process. Two courses of action have been suggested for use. One is to attempt to measure an imputed rent and depreciation to the capital stock by making analogies with respect to what amount of rent the capital item would yield if it were being employed in the private sector of the economy. This rent will include interest and depreciation. But this procedure is subject to a great deal of arbitrariness and uncertainty. Thus, a great deal of judgment is involved in adjusting the estimated shadow prices so that they more closely reflect the true opportunity costs.[5]

An alternative technique for estimating the rate of capital use lies in employing the "capital recovery factor" (CRF). The application of this technique automatically accounts for both interest and depreciation.

The capital recovery factor is that factor which "...when multiplied by the present value of capital costs, is the level [average] end-of-year annual amount over the life of the project necessary to pay interest on and recover the capital costs in full."

The formula is as follows:

$$c = C_0 i(1+i)^n \ / \ (1+i)^n - 1,$$

where c is the capital recovery factor (annual capital cost); C_0 is the present value of capital in use; i is the social opportunity cost rate of capital or investment funds; and n is the number of years over which benefits (of the capital in question) are returned, that is, the project life. In some respects this technique is no less arbitrary than that which imputes rent and depreciation. Apart from the problem of establishing the present value of the capital in use, essentially arbitrary judgments must be made with respect to the values of n and i.

Additional problems exist with the use of this technique. The first is that the CRF does not necessarily indicate the amount of capital used in any given year. It only states the level annual amount needed to recoup the principal and social opportunity cost, that is, interest, given the project life. Second, more than one cohort of students may utilize a given capital item during the life of that item. Here again is the familiar joint cost problem.

Site Costs. Site costs and capital improvements to sites are affected by the joint cost problem, unless, of course, a given site or site improvement is uniquely related to a given output. The site itself is indestructible in most cases since the productivity of the site is not reduced by its use by students. However, the site does have an economic cost since it is productive. An interest charge to estimate the social opportunity cost should be paid, but not a depreciation charge, since, from society's standpoint, the site does not depreciate. This cost should be covered by the benefits gained from the educational process; otherwise, more efficient uses for the site may be foregone, resulting in a loss to total welfare. However, these site costs cannot be sensibly prorated among different educational cohorts.

A serious problem with estimating site costs lies in that such costs are inextricably mixed with the costs of capital improvements to the sites. It is impossible to determine how much of the purchase price of a site is a function of the value of pure rent and how much of it is due to the site improvement.

Cost from Nonschool System Support. Care must be taken to ascertain whether or not the various school systems are subsidized by any branch of the local or state government. Such cost items must be included into total costs. While the school systems of two cities of this study, cities A and C below, do not receive any substantial support from the cities in which they are located, city B receives one-third of its support in this fashion. For this and other reasons the cost data for this school system cannot be used effectively.

Imputation of Indirect Taxes. Some economists argue that adjustments should be made for the fact that school systems do not pay indirect or real estate taxes. Hence a school dollar commands more resources in the market place than does a dollar spent by a private individual.[6] Thus, true social opportunity costs are understated.

However, one can argue equally well that the output from a dollar spent by a school is understated relative to the output from a dollar spent by individuals and firms because normal profits are not charged by the school administration. Thus, adjustments must be made for a downward bias in benefits as well as a downward bias in costs. The present study does neither.

C. Cost Estimations: Three Cities.

Two questions are of interest in an economic analysis of vocational-technical education. First, what are the differences in cost, especially marginal cost, between the vocational-technical curriculum and the different curricula of the comprehensive high school? Second, within the vocational-technical curriculum, what are the relative net economic benefits among the various vocational skill areas? Two sets of cost data are needed to answer these questions.

Vocational Curriculum Costs. Numerous data problems present themselves when comparing the vocational-technical curriculum of the vocational-technical senior high school against the college preparatory, general, and other curricula which are outputs of the comprehensive senior high school.[7]

Inclusions and omissions of cost items such as maintenance cost items are usually uniform within a city school district, but not among school districts. It was not possible to make the necessary adjustments to allow uniform comparison of differential costs among cities.

City A had published cost records but no detailed definitions of cost items included or excluded from the data presented. Nor was it possible to determine these factors for earlier years. Cross-section data by type of senior high school was available through fiscal year 1959-60 but discontinued thereafter. The 1959-60 data is noncomparable in undetermined, but apparently not critical, ways relative to the earlier years. City B had only limited time series data. About one-third of the educational expenditures for this city were aggregated within the city budget. After the 1957-58 school year, average daily attendance by type of senior high school was not kept for this city. Only scattered information on such educational characteristics as class size existed. Cost analysis for this city simply could not be performed. The school district is developing a new program-planning-budgeting system; but due to the budget categories used, no economic analysis of curricula or courses within this school system can be performed.

City C publishes cross-section data by type of senior high school. About 98% of instructional, 75% of operational and 50% of maintenance costs are reported. This city was the only one which had cross-section data by type of school on important structural variables such as median and average class size, or teacher quality. The existence and quality of this data appeared to be in large part the result of the efforts of a few interested at the school district level. Table 1 presents the comparisons of current operating expenditures divided by average daily attendance for the three cities. Because of the fore mentioned dissimilarities in the data, it is not valid to compare these three sets of costs.

TABLE 1

CURRENT OPERATING EXPENSES/AVERAGE DAILY ATTENDANCE FOR SENIOR HIGH SCHOOLS IN CURRENT DOLLARS[a]

	CITY A		CITY B		CITY C		
Fiscal Year	Comprehensive	Vocational-Technical	Comprehensive	Vocational-Technical	Comprehensive	Vocational-Academic[c]	Vocational-Technical
46-47	212	274	217	415	194	119	198
47-48	239	291	280	613	246	151	250
48-49	262	308	318	697	264	155	250
49-50	282	311	309	797	278	166	256
50-51	277	316	320	900	314	200	320
51-52	340	368	349	868	338	187	328
52-53	363	404	351	925	356	175	362
53-54	395	402	378	1,039	363	175	375
54-55	400	471	379	857	380	192	391
55-56	404	489	388	919	405	231	440
56-57	431	528	421	985	420	210	414
57-58	462	587	425	989	429	263	360
58-59	442	607	430[b]	621[b]	444	246	386
59-60	417	702	397[b]	626[b]	498	268	426

Notes:

[a]The use of average daily membership (ADM) or average daily enrollment (ADE) would yield smaller average current cost figures. The reader is cautioned against making unwarranted inter-city or inter-year comparisons of either "quality" or "costliness." These data in themselves imply nothing concerning economic efficiency. These figures are based on current operating expenses and ADA reported separately for each type of senior high school, by city, for the combined ADA of 10th, 11th, and 12th grades.

[b]These figures are based on estimated ADA.

[c]Graduates from this type of high school have, as defined in this study, a curriculum major in both the academic and the vocational-technical curriculum.

Estimated Cost Functions. Equations (1) and (2) describe the statistical functions of total and average current costs. These functions are:

$$(1) \quad TC = A_0 + A_1X_1 + A_2X_2 + A_3X_2^2 + A_4X_3 + A_5X_3^2 + U_1$$

$$(2) \quad AC = B_0 + B_1X_1 + B_2X_2 + B_3X_2^2 + B_4X_3 + B_5X_3^2 + U_2$$

The variables used are defined as follows:

TC = Total current expenditures in dollars.

AC = Average current expenditures in dollars.

X_1 = 1 for vocational-technical schools.

= 0 otherwise, i.e., comprehensive schools.

X_2 = Average daily attendance (ADA, comprehensive schools).

X_3 = Average daily attendance (ADA, vocational-technical schools).

X_4 = Average teacher salaries (total teacher salaries divided by the number of teaching teachers) in dollars.

X_5 = Student-teacher ratio: average daily attendance divided by the number of teachers for city A. Average class size for city C.

U_1, U_2 = Random disturbance terms.

X_1 is introduced to provide different intercepts for comprehensive and vocational-technical schools. Average daily attendance squared is introduced to account for the nonlinear nature of the cost functions. The cost functions are based on the school as the unit of observation. The functions are estimated for the 1956-60 fiscal years.

Equations (1) and (2) amount to implying a state of homogeneity for the quality of education. This assumption, however, is not quite realistic since the quality of education does vary from one school to another. This difference in quality may be assumed to be associated with the costs of instruction. It is therefore useful to modify Equations (1) and (2) to allow for quality differentials in instruction. The concept of the quality of education, however, is an abstract one. It is difficult if not impossible to measure. It is argued, however, that class size and teacher salaries reflect, in part, the quality of education. The reasoning behind this argument is that in a school with relatively small-sized classes, a teacher can devote a relatively large amount of attention to each student. Furthermore, the importance of teacher quality to educational quality is beyond debate. It is assumed that the level of salaries reflects the quality of teachers. This argument is based on the fact that salary level depends on merit, experience and education received by teachers. Also, in a competitive labor market, a teacher's salary may reflect his productivity.[8]

For city A, the estimated results for total current costs are as follows:

$$(1) \quad TC = \underset{(61,678)}{145,624} - \underset{(97,946)}{266,472X_1} + \underset{(68.45)}{284.09X_2} + \underset{(.0180)}{.0062X_2^2} + \underset{(139.41)}{686.88X_3} - \underset{(.0577)}{.0782X_3^2} + \underset{(30,740)}{30,018\ (1957)} + \underset{(30,154)}{102,169\ (1958)} + \underset{(30,196)}{95,264\ (1959)} + \underset{(31,075)}{234,486\ (1960)}.$$

$\bar{R}^2 = .89 \qquad SEE = 94,641 \qquad N = 99$

The estimated results for city A for total costs (current plus capital) are as follows:

$$(2) \quad TC' = \underset{(55,262)}{125,315} - \underset{(66,807)}{147,652X_1} + \underset{(60.00)}{368.02X_2} - \underset{(.0156)}{.0125X_2^2} + \underset{(106.61)}{640.16X_3} - \underset{(.0494)}{.0542X_3^2} + \underset{(29,088)}{45,969\ (1957)} + \underset{(28,789)}{63,332\ (1958)} + \underset{(29,157)}{77,840\ (1959)} + \underset{(30,051)}{164,813\ (1960)}.$$

$\bar{R}^2 = .91 \qquad SEE = 93,029 \qquad N = 99$

For city C, the estimated results for total current costs are as follows:

$$(3) \quad TC = \underset{(87,301)}{203,711} + \underset{(63,096)}{17,753X_1} + \underset{(76.60)}{189.83X_2} + \underset{(.0135)}{.0161X_2^2} + \underset{(81.22)}{211.50X_3} + \underset{(.0149)}{.0377X_3^2} - \underset{(35,845)}{19.484\ (1957)} + \underset{(35,180)}{13,947\ (1958)} + \underset{(33,869)}{59,431\ (1959)} + \underset{(33,011)}{151,560\ (1960)}.$$

$\bar{R}^2 = .95 \qquad SEE = 82,676 \qquad N = 70$

Capital costs were not available for city C.

$\bar{R}^2$ is the coefficient of determination adjusted for degrees of freedom; SEE is the standard error of the estimate; N is the number of observations; and the numbers in parentheses below the partial regression coefficients are the standard errors.

Table 2 presents estimations of marginal costs by current expenditures and total expenditures (capital plus current) for an additional student in ADA in city A. For city C, marginal costs are presented only for current expenditures for an additional student in ADA. Marginal costs for the vocational-technical schools are higher than marginal costs for comprehensive schools in both cities. The major reason for differences in marginal costs between the two types of schools lies in differences in teachers salaries and average class size or student-teacher ratio and not in differences in capital costs as one might ordinarily expect. Capital costs do have an effect on marginal costs, however. As Table 2 shows, marginal costs at average ADA for the comprehensive senior high school in city A rise from $304 to $312 when capital costs are included. For vocational-technical schools in city A marginal costs at average ADA rise from 464 to 485.[9]

By differentiating the average cost function with respect to ADA, setting its partial derivative to zero, and solving for the level of output at which average cost is a minimum, one can estimate the optimal scale of operation for two types of senior high school.

The level of ADA at which average cost is minimum is as follows:

	Comprehensive	Vocational-Technical
City A (1)	2,957	2,295
(2)	3,350	1,958
City C	3,191	3,339

If the statistical results derived in this study are reliable, the optimal scale of comprehensive senior high schools is about 3,000. The inclusion of capital costs in city A raises the optimal size of comprehensive senior high schools by approximately 400 students. The small number of observations for the vocational-technical senior high schools (approximately 15) cast some doubt on the degree to which inferences can be made concerning economies of scale for these schools. The inferences made here are also limited by the fact that we are concerned mainly with total current costs. Capital costs, an important component in the economies of resource use, are excluded from the analysis for city C.

Further Qualifications. The study emcompasses only the fiscal years 1956-60. Cost data for the two types of senior high school in the two cities were pooled for these years. The assumption was that the underlying production function for each type of school in each city did not change over this period of time so that the estimated coefficients of the pooled equations give a better representation of the marginal costs than would equations estimated for each separate year. Cost functions based on current costs were used due to the fact that the value of capital in use was so arbitrarily determined and because of the joint cost problem of the capital with respect to cohorts over time.

Other problems with these cost data remain. First, total current costs include expenditures on additions, renovations and repairs to buildings which are in the nature of capital improvements and which vary from year to year. This variation gives rise to differences in the estimated rela-

TABLE 2

MARGINAL CURRENT COSTS BY AVERAGE DAILY ATTENDANCE FOR VOCATIONAL-TECHNICAL AND COMPREHENSIVE SENIOR HIGH SCHOOLS, FISCAL YEARS 1956 THROUGH 1960, CITIES A AND C, IN DOLLARS

	Comprehensive			Vocational-Technical		
Year and City	Avg. ADA	Marginal Costs at Avg. ADA	Marginal Costs by Linear Approximation	Avg. ADA	Marginal Costs at Avg. ADA	Marginal Costs by Linear Approximation
City A (1)	1,917	304	307	1,426	464	504
1956-60 (2)	1,917	312	321	1,426	485	525
City C 1956-60	2,917	270	285	2,316	386	409

Notes:

For city A, row (1) represents marginal costs by current expenditures; row (2) represents marginal costs by current plus capital expenditures.

tionships between yearly cross-sections. Second, the size of sample used in this study is small particularly for the vocational-technical schools. The observations thus do not include schools of all possible sizes. Under such circumstances any addition or deletion of one school could result in a major shift in the slope of the statistical cost functions.

Costs by Vocational Skill. Only city A had usable cost data by vocational skill. For this type of cost data, the very existence of which is dependent on the capriciousness of administrative accounting requirements, only teachers salaries comprise the average variable costs measured. The skill categories are very broad since individual courses had to be aggregated to generate sufficient observations for analysis. This results in a loss of precision in attempts to judge relative costs and benefits among skills. Table 3 indicates the estimated marginal teacher salary costs for an additional student in ADA by skill category.

The marginal costs range from a low of $106 for woodworking to a high of $415 for the food service skill. No marginal cost for the building trade group is evaluated because, for the functions estimated, total teacher salaries were not significantly related to average daily attendance.

D. Cost Estimations: National Data.

The only source of vocational cost data at the national level comes from the Annual Report on Vocational and Technical Education which is compiled from state reports. Neither the Digest of Educational Statistics nor the Statistics of State School Systems provide cost data for vocational education.[10]

The data published in the Annual Report on Vocational and Technical Education have the following shortcomings. First, enrollment data is published which does not standardize attendance by any index such as average daily attendance. Second, although enrollment by classes, grades, broad vocational specialty and states is provided, costs are not provided on the basis of the same breakdown. Federal funds are broken down but not state and local. Also, for the above breakdown federal funds are on an allotment basis and not on an actual expenditure basis, which makes them unusable for cost analysis. The result of these shortcomings is that any attempts to estimate marginal costs by the seven categories shown in Table 4 is confounded by the fact that the enrollment figures represent diverse types of output. As a result of these shortcomings we are estimating expenditure and not cost functions.

The statistical function of total expenditures is as follows:

$$TC = D_0 + D_1Z_1 + D_2Z_1^2 + D_3Z_2 + D_4Z_3 + U_4$$

The variables used are defined as follows:

TC = total costs (current and capital) in dollars.

Z_1 = total enrollment.

Z_2 = time trend for the fiscal years 1949-50 through 1964-65, with the exception of 1951-52, which is missing.

Z_3 = 1 for the 1964-65 fiscal year.

= 0 otherwise, i.e., all other fiscal years.

Table 4 presents the estimated coefficients. These coefficients represent marginal expenditures for the weighted average of the types of enrollment for an additional unit of enrollment and do not represent marginal costs. Marginal costs cannot be estimated with this data since the unit of observation is the state and not a school or school district. The school or school district is the appropriate production unit. A state is simply a political construct and a meaningful production function for the provision of vocational education cannot be specified for it.

TABLE 3

MARGINAL TEACHER SALARY COSTS BY COURSE BY AVERAGE DAILY ATTENDANCE FOR THE POOLED FISCAL YEARS 1961-1967, EXCLUSIVE OF FISCAL YEAR 1966, IN DOLLARS

Course	Avg. ADA	Marginal Costs at Avg. ADA	Marginal Costs by Linear Approximation
Food Service	89	415	247
Mechanics	170	203	194
Woodworking	74	106	116
Clothing and Fabrics	115	144	161
Electric and Electronics	88	155	202
Agriculture and Horticulture	117	267	260
Personal Service	111	248	260

TABLE 4

TOTAL EXPENDITURE FUNCTIONS FOR FEDERALLY SUBSIDIZED VOCATIONAL EDUCATION PROGRAMS FOR THE POOLED FISCAL YEARS 1950-1965, EXCLUSIVE OF FISCAL YEAR 1952, IN DOLLARS

Vocational Program	Z_1	Z_1^2	Z_2	Z_3	I	SEE	$\bar{R}^2$
Total Programs	478**	.00087	359,682**	3,805,419**	-2,748,606**	3,057	.72**
	(35)	(.00092)	(27,832)	(504,683)	(302,057)		
Agriculture	618**	-.00622	50,845**	48,627**	-146,200**	563	.80**
	(21)	(.00186)	(5,115)	(92,795)	(53,753)		
Distribution	295**	-.02301**	16,808**	112,447**	-103,277**	176	.57**
	(14)	(.00179)	(1,617)	(29,168)	(16,368)		
Health	1,627**	.94807**	4,997	121,775**	-59,928	246	.52**
	(223)	(.32407)	(5,946)	(45,327)	(68,319)		
Home Economics	413**	-.00577**	51,799**	-8,946	-391,096**	724	.73**
	(19)	(.00112)	(6,622)	(119,368)	(69,140)		
Office	223	.02372**	---@	---@	439,007	1,292	.89
	(151)	(.00625)			(290,536)		
Technical	1,960**	-.16855**	12,180	508,801	-126,540	1,106	.42**
	(293)	(.04556)	(113,842)	(291,982)	(1,593,582)		
Trades and Industry	820**	-.00858	46,465**	945,493**	-462,819**	1,153	.69**
	(49)	(.00449)	(10,484)	(189,883)	(111,652)		

Notes:

Numbers in parentheses are the standard errors for the respective partial regression coefficients.

I is the intercept. SEE is the standard error of estimate. $\bar{R}^2$ is the coefficient of determination adjusted for degrees of freedom. * Significant at the .05 level of significance.

** Significant at the .01 level of significance. @ Observations exist for the 1964-65 fiscal year only.

The expenditure and enrollment data analyze how individual states differ in terms of their expenditures on vocational education. The expenditure functions are therefore estimated to relate specific types of expenditures to enrollment in a given program. Marginal expenditures at average enrollment were not estimated since the meaning of enrollment reported in the data is not clear.

The estimates in Table 4 show that both the intercepts and slopes of the expenditure functions differ from one program to another. In particular, in programs such as health and office training programs, the marginal costs increase at an increasing rate while marginal expenditures on agriculture, distribution, home economics, technical, and trade and industry programs, increase at a decreasing rate. The difference in marginal expenditures between programs is at first glance quite surprising. But it should be noted that the expenditure data include capital expenditures for buildings and equipment. For programs which are in an early stage of development, these expenditures may constitute such a large proportion of a state's total expenditure on the particular type of programs. Thus, marginal expenditures are quite different among programs but we have no data to discover why they differ.

Refined data are necessary to analyze the factors associated with the above discrepancies. From the viewpoint of efficient allocation of resources, such refined and reliable data are necessary so that cost and production functions of education can be studied in order to gain optimal efficiency in educational production.

E. Conclusions.

If the federal government, and, by extension, our society, wishes to pursue a more rational course with respect to investment in the human agent, then adequate data must be collected based on sound cost accounting principles and guided by agreed upon objectives and definitions of output to measure those objectives. A national census as represented by the Statistics of State School Systems is not workable. It is too costly and too time consuming given the very limited quality and quantity of cost data collected. It would seem to be much more reasonable to establish a statistically representative national sample of school districts and collect detailed cost data based upon the school as the unit of observation. Within the school, costs should be collected both on a broad curriculum and course or skill basis. Finally, the data should be collected by persons skilled in this endeavor rather than left as it now often is, to a clerk or a secretary at the local school level. As matters stand now, only the grossest type of economic analysis of vocational education can be performed.

FOOTNOTES

* This paper is based in part on a larger study of vocational education being undertaken at the Institute for Research on Human Resources. See Jacob J. Kaufman, Teh-wei Hu, Maw Lin Lee, and Ernst W. Stromsdorfer, An Analysis of the Comparative Costs and Benefits of Vocational versus Academic Education in Secondary Schools, U.S. Office of Education, Project No. O.E. 512, Grant OEG-1-6-000512-0817. Mr. Norman Kalber performed the necessary calculations. The research reported herein was performed pursuant to a contract with the Office of Education, U.S. Department of Health, Education, and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Office of Education position or policy.

[1] See, for instance, U.S. Senate Committee on Labor and Public Welfare, Subcommittee on Employment and Manpower, Hearings, The Nation's Manpower Revolution, Part 6, 88th Congress, 1st Session, October 16, 17, 18, 22, 23, 24, and 25, 1963.

[2] See Jacob J. Kaufman, Teh-wei Hu, Maw Lin Lee and Ernst W. Stromsdorfer, An Analysis of the Comparative Costs and Benefits of Vocational versus Academic Education in Secondary Schools, Preliminary Report, U.S. Office of Education, O.E. 512, October 1967.

[3] See, for instance, Jack Hirshliefer, et al., Water Supply: Economics, Technology and Policy, (Chicago: University of Chicago Press, 1960), pp. 93-94, and Roland McKean, Efficiency in Government Through Systems Analysis, (New York: John Wiley and Sons, 1958), pp. 44-46.

[4] See U.S. Department of Health, Education, and Welfare, Office of Education, Financial Accounting for Local and State School Systems, OE-22017, Records and Reports Series: Handbook II, (Washington: U.S. Government Printing Office, 1965), Chapter 8, "Prorating Expenditures." Other needless confusion and error exists in this document. For instance, the category of Fixed Costs actually contains variable cost items as social security payments and insurance payments. These are more properly allocated to their specific categories such as Administration and Instruction.

[5] For a general discussion of the problem of imputing opportunity costs to resources employed in the public sector see Roland N. McKean, "The Use of Shadow Prices," in Samuel B. Chase, Jr., Editor, Problems in Public Expenditure Analysis, Studies of Government Finance, (The Brookings Institution: Washington, D.C., 1968).

[6] See Fritz Machlup, The Production and Distribution of Knowledge in the United States, (Princeton: Princeton University Press, 1962), pp. 101-102.

[7] These curricula are defined in Kaufman, et al., op. cit., Appendix I.

[8] By assuming that teachers salaries and class size are proxy variables for the quality of education, the total current cost function was modified as follows:

$$TC = C_0 + C_1X_1 + C_2X_2 + C_3X_2^2 + C_4X_3 + C_5X_3^2 + C_6X_4 + C_7X_5 + U_3.$$

The result was to raise the estimated marginal costs somewhat. It appeared that the class size variable overadjusted for economies of scale. The estimated equations are reported in Jacob J. Kaufman, et al., A Cost-Effectiveness Study of Vocational Education, October, 1968. (Forthcoming)

[9] Capital costs are based on historical costs inflated by the Boeckh construction cost index. An n of 60 and an i of 10% are assumed in using the capital recovery factor. For comprehensive senior high schools in city A current costs average $797,910 over the 1956-60 period while capital costs average $57,340 over the period. For vocational-technical schools in city A, the respective costs are $668,810 and $56,000.

[10] See, for instance, U.S. Office of Education, Vocational and Technical Education, Annual Report, Fiscal Year 1965, OE-80008-65, (Washington: USGPO); U.S. Office of Education, Statistics of State School Systems, 1963-64, OE-20020-64, (Washington: USGPO); and, U.S. Office of Education, Digest of Educational Statistics, 1966, OE-10024-66, (Washington: USGPO).

DISCUSSION

by Morgan V. Lewis, The Pennsylvania State University

My remarks are from the perspective of an industrial psychologist who has had the opportunity to study vocational education for the past four years. I have no particular expertise in either economics or statistics, but from my own particular perspective I would like to offer some personal observations about vocational education and then a few specific comments on the papers.

I want to warn you that my remarks arise from a personal judgment--I believe that more occupational-oriented education is necessary in the public schools. I am not advocating the expansion of the type of vocational education that is now generally offered but an expanded concept of occupational familiarization that could bring interest and meaning to the educational experiences of a majority of students.

I believe that the majority of students--something over 50%, but don't ask me to cite a specific figure--are being largely ignored in our secondary schools. Approximately one-third of high school students have the ability and interest to benefit from a college education. Another five to ten percent have a specific occupational goal they can specify and work towards at the secondary level. The remainder need not specific skill training so much as an opportunity to acquire basic work habits and general skills that can be used in a variety of occupations. Even more than this, they need exposure to an education experience that is stimulating rather than stultifying. And this is what I believe an occupational orientation can bring. Youngsters who are bored by the verbal abstractions that dominate the traditional high school course can find topics of intrinsic interest in occupational oriented subject matter that is more congruent with their style of learning, that provides opportunities for achievement, and that has relevance to their lives after they leave school.

It is from this personal position that I approach the question of vocational education statistics. If education is to be guided along the lines I have suggested, what data are needed? First, I think we need some data about the students' reactions to their school experiences--that is their attitudes towards school. I think we would find reflections of discontent and frustration quite prevalent among students in the general curriculum. This is the catchall curriculum for those students who either are not able to or do not wish to enter the college preparatory or traditional vocational curriculums. It is these students who are most cheated in high school.

The attitudinal data could also be helpful in the debate as to the proper setting for vocational education which centers on the comprehensive high school versus the separate vocational-technical school. The reactions of the students, both during their school years and after, should be considered.

Second, I think we need data on the degree to which the students change in the maturity of their vocational considerations during their school years. Vocational maturity is a relatively new concept in the study of career development and career choice but I think the work of Donald Super and his students, especially John Crites, has brought it to the stage that it can be of direct help to the practicing high school counselor.

Which brings me to my third data need, which is more data on the amount and quality of vocational guidance that students receive. Our lack of data is woeful in this area, but it only reflects the woeful lack of vocational guidance in our high schools. The time of most high school counselors tends to be spent in course scheduling and helping college preparatory students to enter college. The work-oriented students are too often ignored. If this condition is to be changed data on counselor activities is obviously essential.

As my last personal point, I think the definition of training-related placement must be sharpened. Most reports of vocational educators state that three-fourths or more of their students are placed in jobs that are related to their training. Studies by other investigators such as Eninger and Kaufman et al. find much lower figures--usually around one-third of the students in jobs related to their training. I am speaking of T & I (trade and industrial) and technical training not of business or commercial education. This latter area does place high proportions of its graduates. This high placement rate in the business area reflects, I believe, the general point being made in these remarks. Business education which has mainly female students, is fundamentally different than the traditional male vocational program which are the trade and industrial and technical fields. Business education is quite general. The skills the young girl learns on one typewriter can be transferred to almost any other. Stenography, office machine, filing systems, business practices, etc., are quite similar in all offices in all labor markets. The skills the young male learns in his vocational shops are much more specific and limited to certain types of work. The chances that he will find a job using these skills are thus much lower.

An expanded concept of occupational education which teaches more generalized skills and at the same time brings meaning and interest to other educational experiences could help to overcome many of the limitations of the current approach.

Now, to some specific comments on the papers. I would like to comment on Mr. Stromsdorfer's first.

I was impressed by the care with which the various considerations in defining costs were delineated and I have neither the competence nor the information to discuss this aspect. I would, however, like to take exception to Mr. Stromsdorfer's discounting of the possibility of measuring quality in education. As a psychologist, I have to contend, almost as an article of faith, that such pnenomena can be measured. I even suggest that performance on standarized achievement test is one indication of the quality of education. Such measures would be more persuasive if they were put into a "value added" framework that would attempt to measure changes across time as a result of defined educational experiences. And, of course, the concept of vocational maturity referred to above, I would contend is another indication of the quality of education.

I wish to support Mr. Stromsdorfer's recommendation for a nationwide sample administered by a central staff that can assure adequate and comparable data. Approximately a year ago the Carnegie Corporation was exploring the feasibility of suchyearly surveys. They encountered considerable opposition from school administrators, and I do not know at what stage their proposal currently stands.

I was happy to see that Mr. Ullman's paper supported my opening remarks for evaluation beyond economic efficiency. (My remarks incidentally were written before I saw the paper.) While I cannot argue with the general framework of cost-benefit analysis, I would repeat again the point made by Ullman that the benefits to be measured need study and specification.

I also would strongly second Mr. Ullman's proposal for more follow-up information. I think it is regrettable that vocational educators, who spend so much time in follow-up studies, gather so little information from them. And even that which is gathered, such as job placement rates, is suspect.

My major exception to Mr. Ullman's paper is that it is so concerned with vocational education as it is. I would hope the Office of Education would be more concerned with gathering statistics that would point the way to occupational education as it should be.

IX

LABOR FORCE LONGITUDINAL STUDIES

Chairman, JAMES SMITH, Office of Economic Opportunity

A LONGITUDINAL STUDY OF LABOR FORCE BEHAVIOR

Stuart Garfinkle - U. S. Department of Labor

Most human interaction depends on some form of mutual understanding in order to operate effectively. Such understanding requires, at best, a history of relevant interaction involving the same people and similar physical conditions. The effective relationships between husband and wife, parents and children, friends, employer and employee all depend upon knowledge of the prior interactions. Similarly manpower policy, to be most effective, must be based on an understanding of the origins and development of worker characteristics which affect labor force behavior. In 1966, to help develop this understanding, the Department of Labor in conjunction with the Bureau of the Census and Ohio State University launched what we believe is the first large scale longitudinal study of labor force behavior.

Prior to the early 1960's, most large scale studies of labor force behavior relied on surveys in which information was sought on the employment status of individuals at one point in time. Changes in the employment situation of the entire work force were studied to determine the relative well being of men and women workers. Such studies had limited use in studying the cumulative effects of a variety of economic, social and psychological factors on given individuals or on groups of individuals who could be identified as having unique problems or characteristics. The Negro, the person with a criminal record, the undereducated, the partially disabled, all have more or less permanent conditions which limit and block paths to a good livelihood.

With this new longitudinal approach we hope to find out much more about these conditions and their effects on employability. Hopefully, also, policies and programs to improve employability can be built on the implications of such a study.

Dr. Herbert Parnes of Ohio State University is responsible for the development of the study and the analysis of the results. The Bureau of the Census has selected the sample and is conducting the necessary interviews.

Two approaches to the measurement of longitudinal data are being used. Changes are measured retrospectively for many variables by asking questions in a single interview about past events and circumstances. Education of parents, first job, best job, date of marriage are examples of such questions. Other changes are recorded by noting differences in annual re-interviews with the same people. Not only can we record authentic changes in this way but because the interviews are only a year apart, and actual records of change are being compared, we can also ask reasons why such changes took place. Some of the changes measured in this way relate to changing jobs, entering or leaving the labor force, and leaving or re-entering a program of education.

Because of the length of the interviews, because the interviews were to be held with a specific person (not with any household member reporting for that person), because of the mass of tabulations that will have to be completed and finally because the same individuals had to be located each year for a period of 5 years, the costs of the survey were estimated to be very high. In order to reach a compromise between available resources and the need to study the most significant problems, study populations were chosen to represent four separate age-sex groups--men 45 to 59 years of age, women 30-44 years of age, men 14-24 years and a group of women of this same age. Each group consisted of approximately 5,000 respondents, selected in a manner to overrepresent nonwhite population by a 3 to 1 ratio.

At this point it seems necessary to briefly describe a longitudinal study and indicate why the Department of Labor became interested in supporting this particular study. Dr. Parnes who is responsible for the development of the study, has defined a longitudinal study as follows: "A longitudinal population study has two essential characteristics. First, it involves measurement or description of one or more characteristics of the same group of individuals at two or more points in time. Second, it involves analysis of relationships among the characteristics of these individuals at different times or of changes in one or more of their characteristics over time."

As noted earlier, information of a longitudinal nature can be collected by asking retrospective questions in one interview or by a series of periodic interviews which take place at different times. It is also worth noting that collection of data covering a long span of years does not necessarily enable a longitudinal study. It is only if comparisons are made between the characteristics of respondents at different points in time that the study becomes longitudinal.

The present longitudinal study of labor force activity was designed primarily to include 5 or 6 annual interviews covering a period of about 5 years. There are three major advantages of a series of interviews as compared for example with a longitudinal study based solely on retrospective questions. First, certain variables are impossible to measure retrospectively. Attitudes toward a job or toward retirement are usually measured indirectly through a specific series of questions. In such circumstances, the respondent cannot be asked to recall what his answers might have been one year earlier. Furthermore, if measurement of change in attitudes is important, the interviews are best made separately so that changes will not be conditioned by the respondents inability or unwillingness to recall a contradictory attitude.

A second advantage of a series of interviews is that the respondent's ability to recall certain events is often diluted by time. Short spells of unemployment, minor sources of income, and different short term jobs are more likely to be forgotten or partially forgotten when the interview covers a long period than when the interview covers only 1 year. The third major advantage of a series of interviews is that comparisons can be made of the ability to recall facts about employment by comparing actual annual interview results with special questions covering the same five year period.

A rather minor advantage of a series of interviews should also be noted. Some respondents will die during the survey period and to the extent that they have significantly different characteristics from the living respondents, it is necessary to interview them each year in order to collect some facts before they die. Only those who survived the entire 5 years could be interviewed in a one time survey.

Although the survey focusses on four age-sex groups in the population--men 14 to 24, and 45 to 59, women 14 to 24 and 30 to 44, the major variables are similar for all these groups. These variables have been described by Dr. Parnes as static and dynamic. Obviously for static variables, only a one-time survey is needed and the information was sought only in the initial interview. Because only the dynamic variables will be measured each year, the length of the second and subsequent interviews will be shorter. Some static variables are color, date of birth, place of birth and first job. Dynamic variables include current employment status, current marital status, current state of health, and attitudes toward work, retirement as well as general outlook.

Because the major purpose in funding the research is to help develop new manpower policies and evaluate the effects of existing programs, the following listing ofmajor variables to be explained emphasizes employment characteristics.

1. Labor force participation will be measured both for a current week of reference and for weeks worked during the year.

2. Unemployment will be measured as of a current week of reference as well as the number of weeks lost annually because of unemployment.

3. Mobility will be reflected in the survey in at least 3 jobs (1) first job after leaving school, (2) longest job and (3) current job. Information will be sought about each of these jobs on occupational and industrial classification, class of worker, length of service, and reason for leaving. In addition to measures of actual job change, attempts will be made to measure willingness to change jobs based on hypothetical job offers.

A great many variables will be examined to seek reasons for changes in the labor force, unemployment and mobility characteristics but obviously they cannot be comprehensive. These variables are broadly grouped into the following related categories:

1. Formative influences including such items as age, nationality, father's occupation, father's education, residence and family composition at age 15.

2. Skills including present and several previous occupations, formal education and vocational training.

3. Health including a measure of relative health as judged by the respondent and a description of healty problems that limit the kind or amount of work they can do.

4. Labor market information including how the respondents looked for work, and how well they advance in their work careers.

5. Marital and family characteristics including number and ages of dependents, the employment, education and health status of other family members.

6. Financial characteristics including current wage, annual income of respondent, unearned income, total family income, and other assets.

7. Attitudes including whether they would work if their earnings were not needed to live on, what they think is the most important thing about a job, satisfaction with current job, and willingness to change jobs.

8. Environmental variables including size of local labor force, local unemployment level, and industrial diversification of the local labor market.

Other variables will be explored for special problem groups such as propensity to retire for men 45 to 59, academic interests and knowledge of the world of work for youths, and marital and family composition and early work experience of women 30 to 44 years of age.

Some Specific Objectives of the Study

Limitations of resources required that certain groups of employment problems be given priority in the study. A selection was made on the basis of the age-sex composition of the work force. As noted earlier, the four groups were men 14 to 24, men 45 to 59, women 14 to 24 and 30 to 44 years of age. Each of these groups has a particular set of problems which affect their employment characteristics.

Men 45-59 Years of Age

This study population was selected because of the incidence of sustained unemployment and involuntary retirement with advancing age among men in this age group. In addition these men are rapdily approaching the institutional age of retirement and information is needed on the relation between attitudes toward retirement and economic, social and psychological circumstances. A third reason for selecting this age group is that the incidence of early retirement is increasing more rapidly for nonwhite men than for white men. 1/ Needless to say, the Department of Labor is very hopeful that the research will uncover a series of policy implications that could affect manpower programs.

Some significant policy questions are listed here:

1. How should training and retraining programs for older men be modified to take into account their education, job experience and physical and other limitations?

2. How effectively can job redesign programs cope with problems of involuntary early retirement?

3. What are the best channels to reach and guide the older job seekers when they lose a job?

4. Is it better to balance local manpower supply and demand situations by encouraging the retraining of men for jobs in their own community when their jobs run out or should men be trained in new skills in order to assist them to migrate and quickly adjust to a new job in other areas?

5. What kinds of institutional barriers can be readily overcome by manpower programs and how can they best be overcome?

6. What kinds of programs can be developed to help prepare older men for retirement and how can these programs be best implemented?

7. What kinds of post retirement employment programs can be developed suited to the economic and social needs of older men and also adapted to their past skills and present capabilities?

Young men and women 14-24 years old

Both of these groups--a separate sample of men and women was selected--were included because of great concern with the high unemployment rates of young people and their obvious inability to make a smooth transition from school to work. Increasing emphasis on the need for more education for workers and job seekers has brought with it recognition that school dropouts are becoming more disadvantaged in the job market. In our city slums, many young men do not even enter the labor force until long after they leave school. Job changing is at a phenominally high rate among young men and women indicating possibly that they may have had inadequate preparation for the world of work, that their aspirations were unrealistic, that job counseling and placement were woefully inadequate or that employers policies toward young workers may need to be revamped for today's world. Some implications for policies and programs that may be developed from this study are phrased in terms of the following questions:

1. What programs can be developed to reduce undesirable aspects of job mobility of young workers?

2. How can the job finding machinery be improved to help young men and women both unemployed and employed find jobs suited to their needs and abilities?

3. What modifications are needed in educational, vocational and training programs to help young men fit more smoothly into the work force?

4. What programs are needed to help those already on a job to make the most of their opportunities or to properly assess their long range opportunities on a job? How can those with limited opportunities be helped to find better work without their having to quit their present job?

5. How can urban slum youths and others with serious employment problems be encouraged to seek and helped to find good jobs? What are the best means of guiding and counseling these persons?

Women 30 to 44 years old

Women in this age group represent three different types of employment problems. The first is that significant numbers are the sole support of their families; for many others it is only by supplementing their husband's earnings that they are able to help their family rise above poverty income levels.

The second type of employment problem is one that relates to severe occupational shortages in such occupations as teachers, nurses and a few professional occupations. Many women with critically needed skills are unable to enter the work force because of a lack of knowledge of specific job opportunities, because of a need for updating of somewhat rusty work skills, or because of inadequate child care facilities available to them.

The third and most general type of problem probably affects more women than the other two more serious problems. How do women with reduced housekeeping and child care work loads find a useful and satisfying outlet for their energies and a challenge to their intelligence. Obviously it will be difficult to develop policies and programs directed to problems as broad as the above. Many of the possible program implications have already been noted in the discussions of younger men and women. Here we are looking for means of accomplishing some of the following specific objectives:

1. What kinds of adjustments in working conditions can be developed for implementation by either public or private programs that will enable women to do paid work without sacrificing family and home obligations?

2. What kinds of training programs can be most useful to women with prior education and work experience to facilitate their reentry into the work force?

3. What kinds of training and placement services can be provided to unskilled women workers to help them to earn a wage which is significantly greater than the offsetting costs of working.

Some unresolved methodological matters

Initially the planning for this study involved a series of 5 or 6 interviews covering a period of 5 years for each interviewed group. Current budget limitations have brought about some modifications in these plans. Present plans are to use a brief mailed schedule in at least one survey for the two older groups in the sample. Other changes may be made but it is expected that only minor modifications will be made until after a great deal of information is obtained and analyzed from the early surveys.

There are a number of technical and procedural problems which will soon have to be resolved regarding the future of the survey. In addition, there are a number of opportunities to add new dimensions to the study which must be considered. Some of the unresolved problems are listed below:

1. What aspects of the study should be continued beyond the initial five year period and what changes in focus will be needed after five years?

2. Should some panels or perhaps all panels be continued beyond the original 5 year period?

3. Should new groups of persons be selected for longitudinal analysis?

4. Should the longitudinal studies be restricted to special problem groups to the exclusion of generally representative population groups?

5. Should the panels be supplemented each year and if so how should the supplementation be made?

One major, but as yet unrealized opportunity lies in the use of school records for the respondents as well as information on the school curricula for the purpose of evaluating intrinsic ability as well as the quality of education received.

Answers to these issues will undoubtedly depend on the usefulness of the results of the initial surveys, how well the surveys provide better answers to manpower problems and, of course future budgets. In addition, new employment related problems may emerge which will be most effectively studied from a longitudinal point of view.

I have described briefly the plans and objectives for this study but as yet have made no mention of progress or results. Older men have been surveyed three times and women 30 to 44 years old twice with one survey of each of these groups by mail. Young men have been surveyed twice and young women once. Analysis of the results of the first surveys of both groups of men will be available this year and subsequent reports on this study should be appearing regularly in the future as more work is completed.

We in the Department of Labor are particularly anxious to see the results of the longitudinal aspects of the study which should begin to become available some time during the next year.

1/ In order to obtain more valid information on the employment characteristics of nonwhite persons who represent only about one-tenth of the sample, a 3 to 1 oversample of nonwhites was built into the survey.

RECENT CENSUS BUREAU EXPERIENCES WITH LONGITUDINAL SURVEYS

George E. Hall, Bureau of the Census

In recent years there has been increased interest among social scientists in the use of longitudinal analysis to determine change and causality in various socio-economic phenomena. The longitudinal approach which takes either the form of a follow-up survey or panel survey 1/ is not new. The Census Bureau has been conducting follow-up studies for a number of years. These studies usually consist of a sample of persons whose characteristics were identified from information provided on one of the Bureau's regular panel studies such as the Current Population Survey.

It is only within recent years however, that the Bureau has become involved in longitudinal panel studies. Currently there is a survey of Medicare recipients in which the members of the panel are interviewed monthly for a period of 15 months. Primarily, the purpose of this survey is to build up estimates of medical care costs. A Retirement History Survey 2/ which will interview a panel of married couples and unmarried individuals, biennially for at least a ten-year period is now in its developmental stage. Both of these surveys are being conducted for the Social Security Administration.

The Bureau's most relevant recent experience has been gained through the National Longitudinal Surveys of labor force experience.

These surveys sponsored by the Manpower Administration of the U. S. Department of Labor, were initiated in 1966. The content and analysis are the joint responsibility of the Manpower Administration and The Center for Human Resources Research at The Ohio State University; the sample design, collection and processing of the data are the responsibility of the Bureau of the Census.

There are four separate longitudinal surveys, each with a focus on a different cohort; males 45 to 59 years of age, females 30 to 44 years, males 14 to 24 years and females 14 to 24 years. The design provides for six annual surveys of each group bounding a five-year period. The initial interviews for each cohort were staggered to permit adequate development of questionnaires, procedures, etc. Data collection for the first survey, Men 45-59 years of age, began in late May 1966; the second, Males 14-24 years, in October 1966; the third, Women 30-44 years, in the spring of 1967; and the last, Females 14-24 years, in January 1968.

The original design calling for an annual personal interview was modified on the basis of experience gained in the early surveys. For the older cohorts, it was decided to move from an annual to a biennial personal interview. A mail survey was substituted in the intervening year to provide current labor force status, and work experience in the preceding year. These changes were dictated by cost considerations and the knowledge that labor force changes for older adults would not be very rapid. Sample persons who do not respond to the mail inquiry are visited by an interviewer to complete the limited questionnaire. Because of the more rapid changes expected from the 14-24 groups, they will continue to be visited annually to obtain more current information.

The National Longitudinal Surveys utilize a multistage area probability sample located in 235 sample areas (PSU's), 3/ which represents the civilian noninstitutional population of the United States within each of the age-sex groups. One of the survey requirements was to provide reliable statistics for nonwhites in each of the cohorts. The sample was therefore stratified in such a way as to be able to increase by a factor of three the probability of selection of nonwhites. The sample was selected to provide a number of households that would yield approximately 5,500 sample persons in each of the initial surveys with approximately 1,833 nonwhites and 3,666 whites. When this requirement was examined in light of the expected number of persons in each age-sex-color group it was found that approximately 42,000 households would be required in order to find the requisite number of nonwhite males in the 45 to 59 age group.

The initial screening interview took place in March and April 1966. During this initial interview a complete household roster was completed in each occupied unit. The name of each household member, relationship to household head, marital status, date of birth, age, race and sex were obtained. These Record Cards were subjected to a clerical review and a "punch card" prepared for each 45 to 59 year old male. From these punch cards four lists of persons were constructed; whites in the white stratum, nonwhites in the white stratum, whites in the nonwhite stratum and nonwhites in the nonwhite stratum. Persons were designated for inclusion in the final sample by simple systematic selection within each list.

The original plan called for using the initial screening to select the sample for all sample groups (only three groups were contemplated at that time, females 14-24 were added later). On reflection it was decided to rescreen the sample in the fall of 1966 prior to the first interview of males 14-24. Males in the upper part of that age group are the most mobile group in the entire population and a seven-month delay between the initial screening and the first interview seemed to invite problems.

To increase efficiency, it was decided to stratify the sample for the rescreening by the presence or absence of a 14 to 24 year old male in the household. The probability is great that a household which contained a 14 to 24 year old in March will also have one in September. However we had to insure that the sample also represented persons who had moved into sample households in the intervening period, so that a sample of addresses which had no 14 to 24 year old males was also included in the screening operation.

This phase of the screening began in early September 1966. Since a telephone number had been recorded for most households at the time of the initial interview, every attempt was made to complete the short screening interview by telephone. These households yielded a total of 5,789 males 14 to 24 years of age and 5,446 women 30 to 44.

The subsequent decision to add a sample of females 14 to 24 years of age again raised the question of rescreening since that survey would not go into the field until late January 1968, seventeen months after the beginning of the previous screening. However the cost of the screening operation itself and the cost of supplementing the original sampling frame with a sample of new construction made it obvious that a better allocation of resources would be to designate the individuals to be surveyed from the previously screened sample. Every effort would have to be expended to find each person selected from this source.

Response rates in both surveys of women tend to bear out the judgment not to rescreen. In the initial older woman's survey the response rate was 94.3 percent; in the younger woman's survey the rate was 94.5 percent. These compare with initial rates of 91.1 percent and 91.7 percent for older men and young men respectively.

The consequences of noninterviews in a longitudinal survey are considerably more serious than in a cross-sectional survey, because of the problem of compounding over time. Therefore, a prime consideration in this series of surveys has been the minimization of noninterviews.

A number of strategies have been adopted to achieve this end. First, it was decided to allow considerably more time in the field than in the usual Census survey. For example in the longitudinal surveys the interviewing typically is permitted to extend for a period of two months although every effort is made to complete the bulk of the work in the first four weeks. This contrasts with the seven to ten days needed to complete the monthly Current Population Survey. Although this leads to inefficiency and higher cost, the value of the completed interview overshadows the negative aspects. For example, in the survey of Females 14 to 24 interviewers were allowed to hold questionnaires from the late January start, until early April when some college students, who perhaps could not have been interviewed otherwise, would be home for spring vacation.

In all of the surveys persons who refuse to be interviewed are sent explanatory letters from a regional office in an attempt to solicit cooperation. The "Refusal" is then visited by a regular regional office supervisor who tries to obtain the interview. Only after such measures fail is a sample person designated as confirmed refusal.

When the final returns are received in Washington, the noninterview rates for each PSU are examined and the actual questionnaires containing a history of the interview attempts are examined for any PSU which seemed to have an unusually high noninterview rate. If on the basis of this examination it is felt that there is a reasonable chance to reduce the rate, the cases were returned to the field for another try. For example in the first survey of Young Men 98 questionnaires were returned to the field. This procedure has not seemed necessary in the more recent surveys.

Traditionally it is felt that one of the most serious interview problems in a longitudinal survey is the problem of finding persons who have moved. During the first detailed interview each respondent was asked for the names, addresses and phone numbers of two persons who would know where he was if he moved. For subsequent surveys this "anchor" provided a reliable means of locating movers.

Names of family and friends were not sought at the time of the screening, so that persons who moved between the screening phase and the initial interview had to be traced by other means. In general, the interviewers used great inventiveness in finding the current address of persons included in the survey. In addition to talking with neighbors and the post office, interviewers checked school records, utility company records, the police and even domestic relations courts when they had information from neighbors or relatives pointing to marital discord.

The 1967 follow-up of males 45 to 59 years of age provided the first opportunity to examine the Bureau's ability to locate people after some passage of time. Of the 5,030 cases assigned in that survey, 97 cases or 1.9 percent were not interviewed because they could not be located. Refusals on the other hand (2.1 percent) were somewhat more of a problem than the interviewers inability to locate the sample person. Only 0.4 percent were not interviewed for other reasons, such as being away, in institutions, etc. for the entire interview period. In addition to the noninterviews 60 respondents died between the first and second interview.

Because of their higher mobility, the young men's survey was thought to provide more of a test of the interviewers ability to find respondents. However there was a smaller percentage of young males lost for this reason, out of 5,234 assigned only 90 or 1.4 percent were not located. The refusal rate was also lower, down to 1.2 percent. Twenty-two respondents from this survey died between the first and second interview.

It appears that when the field operation of a longitudinal survey is reasonably administered the cost is not markedly higher than a retrospective survey of similar magnitude and complexity. Of course this can be misleading. In a retrospective survey all information about the various points in time will be collected at one time, in a longitudinal survey the same information would be collected through a series of observations. Thus for the same information the cost could be higher by as much as a factor equal to the number of observations.

Although it was found that the field costs per observation were not higher than we would expect for a similar non-longitudinal survey, the same statement cannot be made for processing costs. A number of factors tend to increase processing and programming costs. The most obvious one relates to inconsistancies which inevitably appear when data from one observation are combined with that taken from a different observation. It is also unreasonable to establish allocation routines to account for missing data, yet nonresponses to individual questions, like complete noninterviews, have a more serious impact on a longitudinal series because of the cumulative affect. Obviously information allocated at the time of one observation would probably result in an inconsistancy when compared with data from a different observation. Thus considerably more care must be lavished on the processing of longitudinal questionnaires.

A number of problems still remain to be solved in these surveys. Perhaps the largest single problem remaining involves efficient methods of handling the vast amounts of data which are becoming available for each respondent.

One of the methodological problems which remains partially unresolved is the question of including persons in subsequent surveys, who were not eligible by definition for inclusion in the initial survey, for example young men who were in the Armed Forces in October 1966. The sample, while representative of the civilian noninstitutional population at the time of selection, is limited in its ability to describe the same cohort at the conclusion of the survey series. There is no difficulty with the mechanics of selecting samples of persons who were ineligible at the initial period but are eligible now. There is however a considerable cost involved and some difficulty in merging data from these persons into the general analytical framework.

1/ Bruce K. Eckland "Retrieving Mobile Cases in Longitudinal Surveys" Public Opinion Quarterly Vol. 32 No. 1 Spring 1968, pp. 52.

2/ See "A Retirement History Study" by Lola Irelan and Joseph Steinberg elsewhere in these proceedings.

3/ The basic sample design is similar to that used for the Current Population Survey which is described in Technical Paper No. 7 "The Current Population Survey, a Report on Methodology" U. S. Bureau of the Census, Washington, D. C., 1963.

A RETIREMENT HISTORY STUDY

Lola M. Irelan and Joseph Steinberg

Social Security Administration

The Social Security Administration will commence, in the spring of 1969, a ten year study of the American retirement process. Data will be collected repeatedly in biennial surveys from a sample of non-institutionalized respondents, between 58 and 63 years of age, who will number about 13,000 to start with. The result will be a unique body of data on actual changes in the living situations of people as they approach retirement, retire, and adjust to retirement.

The extended, diachronic character of the study results from the realization that a one-time, cross-sectional survey could not produce knowledge about processes. Comparisons among various age groups at the same point in time do not inform about the paths by which separate groups arrived at their current states. This is a point which deserves some emphasis, since the major part of our available knowledge of aging--in particular, of retirement and its effects --has resulted from just such one-time studies.

An important problem in deducing notions about human development from cross-sectional observations arises from failure to study and allow for cohort differences. Individuals, and groups of individuals, have characteristic auto correlated experiences which result from the happenstance of birth at particular points in economic and social history. Leonard Cain has illustrated this by a demographic comparison of the generation of Americans born between 1890 and 1899 with that of 1900 to 1909. The older cohort was more poorly educated, had higher fertility rates, worked longer hours for lower pay, probably suffered more unemployment, and was more likely to be foreign-born. The events of history, the social and economic upheavals through which the country is still passing, have meant different things to these two groups. The younger had no war to fight. It filled instead, more frequently, the well paid defense jobs of the second World War and has enjoyed the country's longest period of virtually uninterrupted prosperity. 1/ Casual comparison of the two groups would undoubtedly produce faulty conclusions about the effects of growing older.

The flaws of cross-sectional studies could be especially troublesome in program evaluation, usually one purpose of Social Security Administration research. Suppose, for example, an enriched social insurance system made it possible for everyone to retire with no decrement in level of living, and to maintain that level the rest of his life. Suppose, at the same time an improvement over some years in everyone's level of living. People well past retirement would be living as they had at the time of retirement. But cross-sectional studies, comparing different age groups at the same point of time, would imply, deceptively, a retirement connected decline in living conditions. 2/

The expense and difficulty of conducting research through time, on the scale necessary for reliable results, 3/ account understandably for the scarcity of examples. Probably the best reported processual study of retirement is that conducted by Cornell researchers in 1952-1957. 4/ Its sample, originally over 4,000, permitted intricate analyses, but was not representative. No women were included. Only industrial and business occupations were sampled. The life-span study of the National Institutes of Health will inevitably cover the retirement process, but its sample is small (600) and limited (white collar and professional men in a limited geographic area). 5/ The Duke University study of aging 6/ started with a volunteer local sample of less than 300. Veterans' Administration is conducting a normative aging study of 1500 men which will include some sociological variables and may throw some light on the retirement process. 7/

Obviously, most of these studies are still ongoing. So the majority of available propositions concerning retirement derive from one-time, cross-sectional observations. The well-known disengagement theory, which postulates a universal tendency for society and the aging individual to withdraw from each other, was based upon a one-time comparison of different successive age groups. 8/ Most published reports concerning sources and ingredients of "successful" retirement rely, similarly, upon either static or retrospective data. 9/ So, too, the more interesting reports of age-associated behavioral change have come from cross-sectional work. 10/

The Social Security Administration's Retirement History Study will add a time dimension to the exploration of a number of existing propositions. In addition new information specific to Social Security program needs will be developed. Concentrated attention, early in the study, will be given to influences on retirement timing and to determinants of retirement patterns. Retirement plans and expectations will be assayed by items on the existence of such plans, their estimated timing, and attitudes toward retirement. In particular, the relation between pre-retirement income, expected post-retirement income, and retirement timing will be examined. The modifying influences of health, anticipated post-retirement needs and resources, employer policies, and significant work history will also be studied. As sample members retire, time-dimensioned data will accrue in several areas where only impressionistic or cross-sectional information is now available. We will learn the actual extent to which mobility is affected by aging and retirement. Changes in consumption practices will be observed, as will changes in activity patterns. The moot question of retirement's effect on health can be studied in context. Later stages of the study, when it is expected that most of the sample will be retired, will include analysis of living patterns.

The accessibility of near relatives has been shown to be an important factor in the lives of older people. 11/ Data will be collected on this area, particularly in its social and economic character. The extent of older people's work life and the factors in various combinations of work and retirement will be analyzed with special interest. Morale and its sources will be studied throughout the ten years of the project.

Primary influences on the design of the survey are implied in its title. As a "retirement" study, it must have a sample of people for whom retirement is meaningful--we shall be studying three 2 year cohorts (58-59, 60-61, and 62-63 year olds) of non-institutionalized men and non-married women. Married women, as individuals, will not be a respondent category. Some salient data on wives of men in the sample will be recorded. When sample males die, surviving wives will continue in the study. As a "history" study, this project will have a large enough original sample so that a validly analyzable number will remain at the end of 10 years. We are starting with about 13,000.

The sample design for the Retirement History Study evolved to satisfy the specifications for analytic needs and after appropriate consideration of the nature of possible sampling resources. At the survey planning stage the analytic specifications are quite general in terms of the wide variety of summaries and comparisons 12/ that are likely to be of interest. The analytic comparisons likely are for 0, 1 variates and different magnitude variates. One objective is the ability to distinguish between a 10 percent change in one cohort and a 15 percent change in another cohort in a given two-year period, insofar as possible. Another is similar ability for the same cohort or different cohorts between pairs of periods. Others considered were to be able to distinguish between relative changes in means of magnitude variates for the same or different cohorts over time. The primary cohorts of interest are not only the two-year groups of all men and non-married women in the target population, but also the two-year groups of married men. The two-year spacing of observations in the surveys and of the cohorts may add a dimension of information additivity, if the parallel observations for different pairs of years of successive age cohorts turn out to be unaffected by the actual passage of events and time.

An initial review of likely sample designs suggested that a multi-stage probability sample would be optimal. Use of an existing statistical survey organization to collect the data seemed desirable and necessary. The exact nature of the multi-stage design was formulated after considering resource alternatives, attendant problems and costs, as well as the substantive needs.

SSA has lists of people in the target population. But these are incomplete. The degree of completeness is a function of the extent to which people in the target population have social security account numbers. This is true for virtually all men but there are material numbers of women who do not have their own numbers. Some consideration was given to the possibility of use of the SSA list of social security number holders as a basic frame for selection of the sample (supplemented by samples of those not in the frame). However, there are no current addresses available to the SSA except for those who are already social security beneficiaries on their own or someone else's wage record. Thus, for the vast bulk of individuals in the target age group that could be sampled from available lists, address information would have had to be compiled. As a minimum, for purposes of use for a given first-stage sample selection, county of residence would be sufficient for each member of the target population--but within the selected sample of primary sampling units full address information would need to be available for a sub-sample. In net, use of this resource would have required full address information for all on the list. A number of experiments were carried out. They suggested that while a good deal of address information could be determined, which would make a substantial portion of the SSA list usable, a complex system of area sampling and weighting in addition would be required to make the SSA list usable, even in part. This multiple frame sampling approach might also have entailed some possible bias because of the need to identify within the area sample those persons who would and those who would not be available finally for sampling from the list frames. The joint effect of cost and potential problems suggested the desirability of considering area sampling as the sole resource. In either event, it was recognized that an area sample was needed, likely, to serve the substantive needs. The size and scope and design of the area sample would be the same under either circumstance.

The possible advantage of the multiple frame sample was seen to be the potential for overcoming, in part, the coverage bias existing for the target population in use of an area sample. The coverage bias that would be reduced would be for those elements in the list frame that are not included in an area sample field process. There is some evidence from a 1960 Census evaluation program matching study 13/ that this is not a minor bias for persons approximating the target age population. Nevertheless, the decision was made to use solely an area sample when both the advantages and the disadvantages were examined.

Area sampling is utilized in many national and local area probability surveys. The Current Population Survey (CPS) of the Census Bureau makes use of area sampling. 14/ As part of the ongoing survey process, the CPS records, among other characteristics, the age, sex, and marital status of all persons who are resident at the sample addresses that are included in the monthly surveys. Preliminary discussions with the Bureau of the Census suggested that this resource, for households no longer included in the ongoing CPS surveys, was one which potentially would be available if the Bureau of the Census were to collect the information for the Retirement History Study.

The final decision on the essence of the sample design for the Retirement History Study is the use of the sample persons in the specified entire target population of the Census Bureau's 449 PSU CPS design. The use of this resource influences the level of the first-stage sample for the Retirement History Study. It was determined that about two and one-half "no longer used (expired)" samples would be required. Operational considerations over the life of the survey suggest the appropriateness of use of the 449 PSU design. The survey mechanism, to be unbiased in the first and subsequent rounds, needs to include not only persons within the sample PSUs but also requires we follow the sample persons to their current addresses at the time of the specific round. The 449 PSU sample design encompasses about 72 percent of the total U.S. population of all ages (according to the 1960 Census). The likelihood, therefore, is that as sample persons in this survey move about, most would move either within their sample PSU or to some location still within the PSU structure of the sample design. Since this target population might be expected to have over 2 percent migrants in an average two-year period, likely less than 1/2 percent of each earlier wave would need to be interviewed outside of the sample PSUs.

The use of the 449 PSU design is desirable also for relatively rapid survey collection without undue training and travel costs. The average workload per interviewer appears to be about what can be accomplished with the amount of available interview time during each planned survey period.

The sample of individuals in the target population will be selected from among the sample in 21 rotation groups of the CPS. The last month in the CPS survey for persons in these 21 rotation groups will be as much as 23 months and as close as 2 months from their first interview in the Retirement History Study. In devising the dimensions of the needed sampling, account had to be taken of several similar factors as will exist in the continued use of the sample. Some of the sample persons will have died since their last month in the CPS sample and some will have moved from the addresses at which they were interviewed in the CPS. Both of these factors over the 10-year interview period of the survey have been taken into account in the determination of the initial sample sizes. Attrition in sample size has been assumed to arise because of death and the possible failure in finding current addresses for some of the movers. The sampling rate assumes that the 1959-61 life table survivorship factors would continue for each age-sex group. It was assumed that these would apply independently of marital status and that the likely joint survivorship of husbands and wives would be the product of the probability of the survivorship of each. Further, it was assumed (hopefully conservatively) that not more than about 5 percent of the surviving sample persons included in a given survey would be noninterviewed in each of the succeeding surveys. This might occur for a variety of reasons. These would be not only cases arising from failure to trace a sample person to a current address but also cases of refusal, not at home on continued calls, or the like. As noted previously, an additional factor affecting the design is the decision to interview husbands and wives as a single unit when both are in the target population. Further affecting the design is the need to follow either if the spouse dies. Thus, the design is affected by the joint age distributions of married couples and the assumptions of survivorship. For purposes of final decision, the joint age distribution of married couples was determined from a CPS subsample. Within each of the two-year age-sex groups, the final sampling plan takes into account each of the differential effects of these factors by age, sex, and marital status.

The substantive results of the study likely will have effects on long-range program planning. In addition, we anticipate that a number of interesting methodological results will flow from this program to add to the stock of knowledge about surveys over time.

FOOTNOTES

1/ Cain, Leonard D., Jr. "Age Status and Generational Phenomena: The New Old People in Contemporary America," The Gerontologist, 7 (1967), 83-92.

2/ K. Warner Schaie and Charles R. Strother have reported a study which documents the different results of cross-sectional and diachronic studies of the same respondents: "The Effect of Time and Cohort Differences on the Interpretation of Age Changes in Cognitive Behavior," paper presented at annual meetings of the American Psychological Association, September 1964.

3/ Reports of a 10-year survey in Great Britain (J.W.B. Douglas and J.M. Blomfield, "The Reliability of Longitudinal Surveys," The Milbank Memorial Fund Quarterly, 34 (1956): 227-251,) emphasize the danger of progressive sample distortion when the sample is not Nationally representative.

4/ Volume 14, No. 2, 1958 of The Journal of Social Issues is devoted to this study.

5/ Stone, J.L., and A.H. Norris, "Activities and Attitudes of Participants in the Baltimore Longitudinal Study," Journal of Gerontology, 21 (1966) 575-580.

6/ Described in various places. See, for example, George Maddox and Carl Eisdorfer, "Some Correlates of Activity and Morale Among the Elderly," Social Forces, 40 (1966) 254-260.

7/ Bell, Benjamin, Charles L. Rose, and Albert Damon, "The Veterans Administration Longitudinal Study of Healthy Aging," The Gerontologist, 6 (1966) 179-184.

8/ Cumming, Elaine, and W.E. Henry, Growing Old, New York: Basic Books, 1961.

9/ See, for example, Charles E. King and William H. Howell, "Role Characteristics of Flexible and Inflexible Retired Persons," Sociology and Social Research, 49 (1965) 153-165, and Aaron Lipman, "Role Conceptions of Couples in Retirement," in Tibbitts and Donahue, eds., Aging Around the World, New York: Columbia University Press, 1962, pp. 475-485.

10/ E.g., Lois R. Dean, "Aging and the Decline of Affect," Journal of Gerontology, 17 (1962) 440-446; S. De Grazia, "The Uses of Time," in R.W. Kleemeier, ed., Aging and Leisure: A Research Perspective into the Meaningful Use of Time, New York: Oxford University Press, 1961, pp. 113-153; and Anthony Lenzer, "Mobility Patterns Among the Aged, 1955-1960," The Gerontologist, 5 (1965) 12-15.

11/ Leopold Rosenmayr, Eva Kockeis, and Albert Kaufmann, "Intergenerational Relations and Living Arrangements in the Course of the Life Cycle, unpublished paper read at the 1963 meetings of the International Congress of Gerontology, Copenhagen.

Peter Townsend, "The Effects of Family Structure on the Likelihood of Admission to an Institution in Old Age: The Application of a General Theory," in Ethel Shanas and Gordon Streib (eds.), Social Structure and the Family: Generational Relations, New York: Prentice-Hall, 1965, pp. 163-187.

Elaine M. Brady and Burton, Gummer, "Aged Applicants and Non-Applicants to a Voluntary Home: An Exploratory Comparison," The Gerontologist, 7 (1967) 234-243.

12/ Sedransk, J., "Designing Some Multi-Factor Analytical Studies," Journal of the American Statistical Association, 60 (1965) 1121-1139.

13/ Marks, Eli S. and Waksberg, Joseph, "Evaluation of Coverage in the 1960 Census of Population through Case-by-Case Checking," American Statistical Association, Proceedings of the Social Statistics Section, 1966, p. 67.

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

15/ When, originally, it had been considered that the target population would include wives, the use of double sampling within the survey of sample persons received active consideration.

DISCUSSION OF LABOR FORCE LONGITUDINAL STUDIES

J. W. Wilson, North Carolina State University

It is obvious from these papers that the two longitudinal studies will provide a great deal of new information on peoples' decisions to enter into, participate regularly, or retire from the labor force. The fact that the samples are relatively large (roughly 13,000 for the retirement study, and 20,000, four groups of 5,000 each, for the labor study) and well designed are encouraging from the standpoint of statistical inference. A great deal of planning and thought have already been invested in the design of the sample. The survey data should be very reliable and allow for powerful statistical tests of competing hypotheses.

The papers, however, have given us little insight on any specific hypotheses to be tested, or generally how the data are to be analyzed. Both longitudinal studies appear to be for the stated purposes of the evaluation of existing policies and the proposal of new policies, but before data can become useful for answering policy questions a thorough analysis must be performed. These papers provide little information regarding the economic or sociological model underlying labor force or retirement behavior. Much theoretical and empirical work has been done on labor force participation, and retirement decisions have been receiving increasing study recently. There are many hypotheses with policy implications that urgently require statistical testing. However, the ability to test many of these hypotheses will be determined partially by the type of data available. It appears that more resources have been applied to the design of the sample than to how the data are to be analyzed. These papers have provided only vague information on the type of data forthcoming, the variables measured, and the relevance of the data to policy questions.

The analysis of such a vast amount of intertemporal data raises some crucial questions. Economists and statisticians have had a great deal of experience in using standard techniques for assisting in the analysis of cross-sectional data and for the analysis of time series data. But, will the standard tools be sufficient to <u>efficiently</u> analyze the longitudinal data with the repeated observations on the same sample group through time? Many hypotheses may be tested with any one of the yearly samples with standard tools, but in order to use the longitudinal data efficiently, it would seem necessary to analyze the intertemporal and cross-sectional data simultaneously. Techniques such as regression methods combining time series and cross-sectional data, and matrix models of interregional mobility will be required for the simultaneous analysis, and, no doubt, new techniques will have to be developed.

Finally, I would urge the Social Security Administration to make these data widely available to Business and University researchers. Many different approaches will be required to thoroughly analyze the data from these important samples.

X

METHODOLOGICAL PROBLEMS OF EVALUATION RESEARCH

Chairman, EDWARD A. SUCHMAN, University of Pittsburgh

STRATEGY AND TACTICS OF EVALUATING A LARGE SCALE MEDICAL CARE PROGRAM

Howard R. Kelman, Columbia University School of Public Health and Administrative Medicine
and
Jack Elinson, Columbia University School of Public Health and Administrative Medicine

The emergence within recent years of a variety of large scale medical care programs which seek to extend traditional patterns of medical care services in new ways, or to drastically modify them, has created urgent challenges for those concerned with evaluation of their efficacy. The magnitude of some of these programs, the complexity of the services offered, and the peculiarities of their origin and development call for modification or new evaluation research strategies.

The material which follows reports some early experiences of an evaluation unit whose activities have been designed to provide continuous appraisal of the impact upon an urban ghetto community of a program through which the resources of a community general hospital center have been merged through affiliation with those of a University Medical Center. Evaluation in this effort has been conceived of as a built-in monitor of the program accomplishments and shortcomings, if you will, with the dual objectives of providing both long term as well as more immediate feedback of programmatic consequences. This evaluation program differs from the more usual types of evaluation studies, at least in its conception, in that appraisals are to be continuous and not one shot.

This report will illustrate some of the methodological issues and problems encountered thus far in developing a program of evaluation activities tailored to suit the specifics of the situation confronting us. First the setting and background will be sketched briefly.

The medical care program we seek to evaluate evolved against a background of growing professional and community concern for the declining levels of care that had become increasingly evident in the municipally operated hospital system of New York City. In an effort to halt this retrogression in care and to upgrade its quality, a program to affiliate the municipally operated institutions with medical schools and voluntary teaching hospitals was undertaken. Under this program, the operation and supervision of the medical and related professional services of the municipal institutions were to be assumed by the medical school or the voluntary teaching hospitals. These agreements, in addition to providing funds to obtain additional staff, equipment and other resources, provided also for necessary programmatic expansion and for related teaching and staff training activities. However, the program we will be reporting on is the only one of some twenty odd such mergers that provides an evaluation component in a general hospital setting as an integral part of its operation.

The history of the development of this vast affiliation effort has not proceeded without controversy as to its effectiveness and costs. Political interests as well as professional concerns, since the inception of the program, have served to keep public and professional attention sensitive to the various activities that have marked the program's development.

The hospital of our concern is a 700 bed physically outmoded institution - soon to be replaced by a modern structure - located in the heart of a Black ghetto community. It services an area containing some 400,000 persons and is the major health services center in the community.

In addition to the kinds of social currents and unrest that prevail in many such communities today, the University itself has for a variety of reasons become a convenient target for political groups in the community who are suspicious of its intentions. From time to time these groups have attempted to arouse community opposition around a number of issues including the health of the community and the operation of "their" hospital by the University Medical Center. It is against the background of these past and continuing social and political forces that the affiliation of the University Medical Center with the community hospital center of our immediate concern has taken place.

The affiliation process itself was accomplished over a period of several years on a department by department basis. More recently it has culminated in an overall institutional contractual agreement, except for one department. The evaluation component of the affiliation is now organized as a separate group within the affiliated community hospital and is linked to the University Medical Center by virtue of its being a function of the University's School of Public Health and Administrative Medicine, which in addition to the evaluation activities, carries responsibility for the operation of certain other services conducted at the community hospital site.

A wide variety of questions have been put forth as being salient to the evaluation unit's attempt to assess the impact of the affiliation program. The questions or issues cluster into two major areas of concern:

1) questions addressed to the community - its health levels, patterns of care, needs for care, and how these needs are met by the multiple providers of care, including the hospital, in the community; and

2) questions centering around the care given by the hospital - its appropriateness and adequacy and its consequences for the

recipients of care.

The first set of questions is being approached through a community based survey of representative households whose health status and related behavior, and changes in status and behavior, are being studied over time. Studies of the character and consequences of care in and around the hospital and its patients address themselves to the second set of questions. The two sets of studies, and questions, are viewed as complementary. In one - the Community is the "target," in the other it is the hospital patient which is the "target." Together, they ultimately will provide information concerning the hospital's programmatic impact.

Another way of defining the relationship between the two study efforts is by viewing the community-based study as providing information about the "denominator," while the hospital based studies will tell us something about certain "numerators."

In developing our evaluation activities we have had to take the following factors into account:

1) Our entry into the hospital affiliation program was at some point after its inception. Thus, we do not have the possibility of a before-after evaluation strategy. Our approach is probably more akin to a "during-during-during" strategy.

2) The identification of the beginning, the midpoint or even the end of the care program is impossible to define in a general hospital which was already an ongoing operation prior to the affiliation of the institution with its counterpart University Medical Center.

3) The Hospital Care program is not a readily definable package of the same order as a drug, a public housing project or even a rehabilitation effort. What we are confronted with is a dynamic, complex, extremely busy hospital in which affiliations have developed over a span of several years with counterpart departments at the medical center on a service by service basis. Thus some services, or sub-programs, in the overall hospital program may be as much as three to six years in the making, others perhaps one year old and still others are on the drawing board or in various states of planning and development.

   An advisory committee, which meets monthly, serves to keep us abreast of program developments and provides also an opportunity for the hospital directors of services to spell out their departmental goals, problems and accomplishments.

4) We have not found it possible to capture or to reconstruct more than fragmentary base line information - relevant to evaluation of programmatic accomplishment - and have had to plan ourselves for the creation and gathering of basic descriptive information relative to the community, the sociodemographic character of the population, health resources and so on. These sorts of information can then be used as base line measures to assess change in denominator and numerator populations in utilization patterns, health levels, and related questions concerning health care in the community.

5) The creation of a reliable body of information at the hospital site useful for evaluation purposes has also had to be undertaken by us. Our need for reliable information relating to the treatment of patients that would also be amenable to electronic data processing merged also with a need on the part of the hospital for systematic utilization review. Currently in its early stages of tryout - after nearly a year or more of development - is a systematic record abstract which contains a large number of items of information concerning the ingredients of treatment. It is meant to be completed on all patients discharged from the hospital. It is anticipated that such information on the nearly twenty thousand discharges per annum will be a valuable source of information for purposes of medical and record audits, or more broadly put professional audits of care and for utilization review. It may also serve other hospital needs for statistical and other reportage of effort. Conceivably follow-up and other kinds of special studies for purposes of evaluation can in the future be designed and executed. It is also anticipated that this activity may have the effect of "beefing up" other record keeping systems in the hospital.

6) Another set of issues we have had to confront has been the differing definitions of evaluation by hospital personnel on a variety of levels, and flowing from this, their differing demands or expectations of evaluation activities.

   One view of the purposes or functions of an evaluation unit sees its role as being akin to that of an inspector general in a military establishment. Evaluation activities defined thus would concern themselves mainly with checking up on the manner and mode of the detailed delivery of services to patients. The evaluation unit would find ways of uncovering the shortcomings of care, bring them to the attention of the providers and check on their change or correction.

   Another conception of evaluation would define its functions as akin to a statistical and research service bureau of the hospital. The function of the evaluation unit in this view would encompass such activities as the gathering and compilation of statistical reports on hospital activities and on the community and its population. Also the unit would provide technical consultation in the design and conduct of research, record keeping and data processing and assume responsibility for providing computer support to ease the problems of administrative bookkeeping.

Still another conception of evaluation defines it as a kind of follow-up service operation. This view held largely by those with direct patient care responsibilities looks to the evaluation unit to provide them with information on the fate of individual patients. A varient of this viewpoint also looks to the unit as a resource for resolving administrative and programmatic problems and questions.

In these several viewpoints there are obvious elements of compatability with our own role definitions and activities as they have developed. There are also obvious pitfalls inherent in each view which we have thus far successfully avoided. To play the role of the inspector general would, we believe, quickly embroil us in entaglements with staff responsible for care. The role of the overseer we feel is not compatible with objective evaluation of care and its consequences. To become a sort of statistical or research service bureau while undoubtedly helpful to the overall operation of the hospital would, we believe, deter us from the task of appraisal of care and relegate us to the status of a bookkeeper or accountant. Finally, we do not define ourselves as being competent to provide clinical and program administrative services even assuming that the potential rivalry of an additional service unit could be avoided.

We have contended with these tugs and pulls through constant repetitive interpretation of our role and through providing limited consultation on research, record keeping, data processing and related topics. Members of our staff also sit as members of ongoing hospital administrative, review, and research committees. We are coming to be viewed as a member of the hospital family - a new and different one perhaps - but still not as outsiders who may threaten exposure.

As was indicated earlier, the overall design of our evaluation activities provides for studies of in-hospital care and its consequences for patients.

The particular approach we have been developing evolved against the background of both general considerations concerning evaluation of hospital care as well as specific features of concern centering around our particular hospital program.

Approaches to the evaluation of hospital or medical care are varied, and include:

1) follow-up or "outcome" studies of selected groups of patients,

2) studies or audits of the quality of medical and surgical care,

3) studies of the structure and functioning of the treatment environment or system

4) experimental program assessments and/or clinical trials, and

5) statistical compilations of hospital or health activities.

It is clear, in considering the multi-functional properties of an acute general hospital, that each of these approaches (and others not cited) has or could have a contribution to make in providing information as to how a hospital functions, and in gauging the effects or impact upon patients of changes in hospital treatment program. It is also clear that each approach, or strategy, though it would be contributory would also be incomplete when measured against the complex task of evaluating a total hospital care program. As yet there is no single, all inclusive, research model or methodology that can be readily applied to the hospital situation we face or to any community general hospital for that matter.

For example, counts of hospital or health activities (e.g., number of admissions, mastoidectomies, visits to clinics, board certified medical doctors provide information concerning program effort or the achievement of program activities. But, they say not very much about patient benefit - effects or outcome. The same is true in part of the "medical audit" approach. Clinical trails tend to be either diagnostic or treatment-specific and require highly rigorous field control conditions.

As we have sought to develop and tailor our own method we have leaned heavily on two approaches which have been successfully employed in prior medical care research. The first is the more or less traditional medical audit whereby judgements of quality or effectiveness of care on selected diagnostic categories of patients are made by some expert or peer group based on information contained in the hospital record. Such studies have employed either standardized preselected criteria or have asked for overall, less rigorously defined ratings of care from expert groups.

In our own work we have broadened the concept of medical audit and are attempting to secure ratings which go beyond the technology of medical or surgical treatment to include nursing as well as social work services.* We have been somewhat arbitrary in deciding to go this far and no further because conceivably all facets of the hospital and how it functions can and do influence patient care. Though we may alter our view at some later point, these seem to us at this time to be the most consistently crucial elements.

We also are endeavoring to increase the information base on which these judgements of care are made by having the members of the assessment team not only review the record, but also see the patient and clinical staff responsible for the patient's care. These reviews of care will be made while the patient is in the hospital and available for study.

In addition to these judgements of care the team is also rating "outcome." Outcome is defined in terms of health and social functioning along a variety of dimensions including clinical course, life threat, self care, ambulation, productive activity and discomfiture to self and to others. The team makes ratings of patients' levels of function on these dimensions at or close to the point of discharge and it reconstructs them for status at admission. In this way we hope to identify change in patient status occurring during the hospital stay.

An additional set of judgements or ratings asked of the team involves judgements of expected outcome, again employing the same criteria. That is to say the team is asked not only to judge patient status at different time points, but also whether the observed change in status is different or similar to that which this patient ought to have achieved were optimum care available and applied. Where a discrepancy occurs between achieved and expected status, the team is asked to identify, if possible, the reasons for the discrepancy. These may include specific factors relating to the content or delivery of care in the hospital, untoward and unforeseen events affecting the patient's clinical course, or factors not intrinsic to care or the system of care.

We have observed thus far two kinds of discrepancies. In the first instance, expected outcome exceeds achieved outcome. Here one would look to see whether and to what extent upgrading in kind or quality of care could or would be beneficial. We have also found instances where achieved outcome exceeds that of expected outcome. Here one could say that the patient's level of functioning was not "disabling" enough for his or her well being.

*These ratings are being made by an assessment team consisting of a physician, nurse and medical social worker.

Where there are no discrepancies between expected and achieved outcome, we would argue that upgrading in kind or quality of care is not necessarily indicated in order to improve function, or if it is provided, benefits should not be sought for in terms of the criteria or outcome measures that we are using. (Upgrading in care may be entirely justified for other reasons of course.)

We are coupling this appraisal of hospital care and its consequences for patients with studies of the patient at three months and also at twelve months post-discharge. Again, employing the same outcome measures, patients functional status will be reassessed. Additional information concerning after care utilization, social and economic circumstances and patients' views of care will also be collected. For the development of this phase of the work we have leaned heavily on the methodology of post-discharge follow-up studies.

Our decision to use the functional status of patients as the yard stick to "measure the results of care is based upon the following considerations:

1) Particularly in a municipal hospital situation, a wide variety of forces - of which the technical competencies or performance of physicians and other professional staff is but one - may affect patient care and influence outcome.

2) Outcome is also likely to vary and to be affected by the patient's past bio-social history and experiences - his potential to respond to treatment - whatever the character of treatment it self is.

3) The patient's post discharge status is also likely to be influenced by still another group of factors - social situational, post discharge services, etc., all of which may influence his clinical and functional state.

To select out of these different matrices of factors one or two treatment variables only for study (e.g.: Medical Doctor's performance, length of stay) means we also need to be able to hypothesize about their presumed consequences for outcome of care. This, to say the least, can be extremely hazardous. It is possibly less hazardous and maybe more relevant to our overall evaluation goals to define and describe changes in the character and status of the patient population over time.

It should also be noted that we do <u>not</u> propose to measure the specific ingredients of care or services (which have been delivered or received by the patient) per se. Nor do we propose to measure the quality of care per se. We do ask that the assessment team address themselves to these questions in so far as it is possible for them to define and analyze their consequences for, or effects upon, the patient's clinical and functional status.

The primary focus then of the judgements of the assessment team is the patient - his status and changes in status during the course of hospitalization and subsequent to it. It is in relation to evident, expected, and predicted, status changes that other judgements concerning patient care practices (clinical administrative, social) and their influence on outcome (patient status) are to be made. Through these means, if our efforts are successful, and coupled with study of the patient following discharge, we hope to trace and link up ultimate changes in patient status to patient care factors in and out of its hospital. It is our expectation that the information produced through these efforts will enable us to address ourselves to the following questions believed salient to an evaluation of the hospital care program.

1) In socio-medical terms, what is the character of the population now receiving hospital care and what is its potential for change?

2) To what extent, and for what kinds of patients, are care needs being met or not, and with what consequences for the patient and others around him?

3) To what extent, if at all, could or would upgrading in kind or amount of hospital care be expected to achieve different "results"?

It is possibly worth while noting here our belief that the validity of the procedures we are developing rests upon the following set of assumptions. The first assumption is that care currently being offered in certain areas is something less than optimum. This is so, either because of an insufficiency in kind or quality or overwhelming demand, or because of a poor potential for favorable response on the part of patients, or some combination of these factors. Secondly, it is the hospital's job to favorably alter the health status of patients where it is possible to do this. In addition we assume that the hospital's job extends beyond responsibility for the care of the patient while he is inside the institution and includes provision for after care where indicated. Finally we believe the dimensions that we have chosen to measure or to determine patient status are salient.

The decision to focus initially on the in-patient hospital stay is not purely an arbitrary one. While ideally in studying medical treatment it would be best to begin at or near the origin of an illness episode, to do this at this point in the development of our overall evaluation program is not feasible in terms of establishing an "at risk" but presumably "well" population. The desirability of having a population which can be defined clinically and dianostically ruled out, again on the basis of feasibility, using the hospital admission as the entry point of study. (A useful approach if one is studying a numerically restricted group of patients with very well defined and established medical conditions.)

Using a discharged population permits a greater latitude of choice and selection of representative patient populations and, also relevant to objective evaluation, obviates the possible bias of observed treatment. Another advantage of a "soon to be" discharged population resides in the background of past experience of studies of quality of care which have focused similarly on the hospital stay of discharged patients.

Just as the origin of the illness episode may not coincide with hospital admission, the termination of an illness episode may not conclude with hospital discharge. In these assessments the discharged population will be studied three months after leaving the hospital and again at 1 year post-discharge. This is planned for several reasons. First, it is desirable and important to relate the outcome of care to estimates of the character of care received in terms of immediate consequences (soon after discharge) and then again after some time lapse (residual effects). Secondly, it is of importance to get the patients' side of things. (events around the need for hospital care) and their views of care and services received as well as other situational factors affecting post-hospital status (after-care services, etc.). Third, a prospective view of patients' function and behavior will permit study of how the illness episode is "concluded" (if in fact it is), how and if other arms of the hospital are employed (Out-Patient Department, Home Care, Emergency Room); how other health and social agencies are called upon by providers of care or by the patient for continuing care; and of equal importance it permits observation of "new" illness episodes which may emerge during the 12 month post-discharge observation period.

It is to be noted that these evaluation efforts go beyond available evaluation studies in two important respects:

1) They extend the observation period well beyond the hospital stay.

2) They seek to "measure" both care and outcome of care and propose to relate them to one another as well as to other factors.

To conclude we might summerize by saying that our efforts thus far might properly be classified as being in the pre-evaluative, baseline data gathering, and methodological developmental stage. In addition to the complexities of the program and the difficulty of defining the relevant evaluative questions we also find ourselves working within a system which is not geared to the demands and requirements of evaluation research. The system is itself being retooled for more efficient, up-to-date and more humane patient care. In this process the role of an evaluation group is subject to varying definitions and strains. To a large extent the pace of its development proceeds at a rate determined largely by the ability of the medical care system to respond to the total demands upon it for patient care in which evaluation is but one relatively minor component.

Our experience to date suggests that built-in, long-term, evaluation of the work of complex large-scale health programs and organizations presents in its aggregate, if not in its specifics, some very special theoretical as well as practical problems, in contrast to more limited, less complex "fewer faceted" service programs. Some of the reasons for this are:

1) The "target" is a community or some other geographically defined population, itself embedded into a larger population matrix.

2) The program objectives cannot be neatly or parsimoniously defined because perhaps, like "sin" the program has always been with us, or if new, some of the goals are global and diffuse and the essential worthwhileness is self-evident.

3) The application of an experimental design is either inappropriate or impossible.

4) The research models we have at hand - crude as these may be when measured against laboratory or other exquisite standards - were derived from the pilot, contrived demonstration or experimental trial "era." While no doubt the need for, and utility of, such endeavors will continue, they do not seem to fit the "new" demands for evaluation that we, for example have encountered. In short, as we have "discovered," there is no single inclusive, well developed methodology that can be readily identified, and applied.

5) As a result of these factors and others, such evaluation programs are likely to be long term and "high risk" in terms of their immediate pay-off, and eventual yield. This has obvious implications, career and otherwise, for evaluation research personnel as well as for the sponsoring or "host" agency. It also has administrative implications with respect to the structual relationships between the evaluation and service programs that are required to assure the achievement of both service and evaluation goals. All of these problems assume different proportions in long term as opposed to single shot evaluation projects.

CONTEXTUAL CONSIDERATIONS IN EVALUATING NARCOTIC ADDICTION CONTROL PROGRAMS

Alan S. Meyer, The Burke Rehabilitation Center
and
Stanley K. Bigman, The Washington Center for Metropolitan Studies

The Changing Context for Addiction Evaluation

Implicit in the task of evaluating any program directed at changing human behavior is the need to describe and understand the social and cultural context in which the program is being carried out. This involves understanding the society's attitudes and policies towards the kind of behavior which the program is trying to change. It also involves understanding the influence of this context -- whether constraining or liberating -- on the problem, on the program, and on its evaluation. It is the purpose of this paper to make explicit some ways in which this context might usefully be considered and to define the kind of evaluator role required for this task.

The problem of narcotic addiction provides a strategic example of the necessity of explicitly considering the cultural context of a problem if efforts at evaluating programs to control the problem are to be meaningful. Addiction is a strategic case study in evaluation for two reasons. First, it is a socio-medical problem which is still viewed and handled in some countries -- and our own has been a prime example -- as basically a criminal problem.

Briefly, the history of opiate addiction in this country has been marked by a dramatic change in the characteristics and behavior of addicts following federal action aimed at transferring treatment of addicts from physicians to police. Major results included the growth of a black market, the development of a criminal subculture among addicts, and the recruitment of new addicts increasingly from minority group youth in urban slums.

In recent years, this predominantly punitive approach to addiction, which is relatively rare in the health field, has been moderated by a growing emphasis on mental health aspects of both etiology and treatment. This has resulted in a highly ambivalent cultural context in this country in which addicts are viewed and treated as if they were both criminal and sick.

The second reason that addiction provides a strategic view of contextual problems in evaluation is that significant changes are taking place in the social and cultural context in which treatment of addiction is being carried out. Public policies, if not popular views, have in the last four years begun to undergo major changes with regard to experimentation in the treatment of addiction. For the first time since the 1920's, when narcotic clinics dispensed drugs in a number of major cities in this country with varying degrees of reported success or failure, physicians have obtained the permission of policing agencies to use maintenance dosages of narcotics in the treatment of addiction.

This clearly represents a major change in public policy. A number of physicians in different parts of the United States have successfully reasserted their right to establish treatment goals and methods in this field. When professionals legitimately engaged in treatment are subject to governmental or other constraints which severely limit their freedom to establish the goals and methods of treatment which they may consider professionally appropriate, then questions can be raised not only with respect to the proper professional posture of treatment staff but also with respect to an appropriate approach to evaluation of such programs.

The Role of Evaluator: Data Analyst or Behavioral Scientist

One view of evaluation holds that the evaluator is a technician responsible only for accepting the program's stated goals and preferred success criteria as givens. His job is to implement the basic evaluative design mandated by the program staff. Such a role has been referred to in the addiction field as a "data analyst."

This role might permit the evaluator to make minor modifications in design through his technical suggestions and skills. However, basic questions regarding the "appropriateness" of the program goals and to some extent even of its methods would generally be considered outside the responsibility of the "data analyst."

We would like to consider an alternative approach which views the evaluator as a behavioral scientist rather than as a data technician. A major part of the behavioral scientist evaluator's responsibility would be to raise basic questions about every aspect of the program, including the social and cultural context in which it is carried out, and about the evaluation design before, during, and after the evaluation study.

The key questions to be raised by the evaluator, as we see them, and which we shall discuss here, are: (1) what is the sponsor's purpose in having an evaluation; (2) what are the implicit as well as the explicit goals of the program; (3) what aspects of the program should be included in the evaluation and what aspects, if any, can be excluded; (4) what should be the criteria of success; (5) what methods of measuring success criteria are scientifically and ethically appropriate; (6) how should the recipients of the program be classified; (7) who should do the evaluating; and finally, (8) what are the constraints on public and private dissemination of findings.

On the basis of the answers to these

questions, it seems to us, behavioral scientists can make a professional judgement as to the nature and extent of their role in the proposed evaluation. A more detailed look at these questions as each applies to evaluation in the addiction field highlights a number of contextual problems which evaluators in other fields might usefully take into account.

What is the Sponsor's Purpose in Having an Evaluation

The purpose that people have in mind for doing an evaluation in the addiction field as in other fields is typically complex. Purposes may include a combination of any of the following: to see how well a program is doing, to document the need to maintain or extend the program, to find out how to improve the program within traditional limits, to learn whether to drop the program, to stimulate fundamental innovative changes, to give an aura of scientific respectability to the program, to delay action in a controversial area, to increase our understanding of the nature of the problem which the program is aimed at solving, to meet an administrative requirement, or to achieve some other purpose.

The original purpose or purposes are not so important as the ones ultimately stated after discussion between the program staff and the behavioral scientist evaluator. The latter may decide on the basis of the answer to this question alone that a meaningful evaluation is not relevant or possible. He may, on the other hand, achieve a modification of purpose which at least allows for the possibility of an evaluation which meets professional standards.

The problem pointed up by much of what has been called evaluation in the addiction field is that the purposes have generally been too narrow to allow either for an increase in theoretical understanding or for a realistic consideration of alternative programs. Thus, the implicit purpose of much addiction program evaluation has been to maintain the status quo in public policy as well as in theoretical understanding. It is of course true that demographic data on addicts and relationships of these data to abstinence rates and other post-discharge behavior of addicts provide increments of data which increase somewhat our general fund of knowledge. The question, however, is whether such increments obtained by studies conducted primarily for fund raising, administrative bookkeeping, or public relations can justify the term scientific evaluation.

What Are the Goals of the Program

To evaluate the effectiveness of a program, it is necessary to know the goals of the program. Evaluators are familiar with the difficulties in accomplishing this seemingly simple task. Goals may contain internal contradictions, cliches, ambiguities, pre-conditions, etc.[1]

In narcotic addiction, program goals have tended to be unclear, implicit, contradictory and shifting for several special reasons. In part, because of our ambivalent view of addicts, and in part, because of the extremely high rate of failure in most programs and the subsequent likelihood of staff frustration, explicit goals tend either to be very broad -- i.e., help control addiction, help addicts, etc. -- or to shift to research or training goals.

Perhaps the major reason for lack of clarity in the goals of these programs, however, is the severe governmental constraint which all such programs have until recently labored under. From the mid-twenties to the mid-sixties, law enforcement agencies directly prescribed the methods, and indirectly the goals, of programs to treat addiction. By preventing any use of narcotics in treating addicts, except as a detoxifying drug, federal policy mandated abstinence from drug use as the universal ingredient in all treatment programs of this chronic condition.

Frequently, programs for treating addicts -- whether in hospital or jail -- have been implicitly aimed at achieving the goal of abstinence through the method of detoxification. In recent years, explicit goals have often included such objectives as rehabilitation, return to productive life, enhanced ego functioning, and longer periods off drugs. So long as detoxification remains the core modality, however, the de facto goal tends to revert to abstinence.

In the last few years, alternative program models have been presented which explicitly classify abstinence, detoxification, individual and group counseling, etc., along with drug maintenance as methods, not as goals or subgoals. The explicit goals, according to this rehabilitation-based model are improved social and psychological functioning and improved physical health within the limits of chronic disability (i.e., without regard to whether the patient or client is on or off drugs).[2] In such a model, explicit and implicit goals tend to merge.

Clarification of program goals must therefore take into account the possible influence of social and cultural contexts on the manner in which goals are stated and adhered to. The goals in turn will determine the criteria of success which can be logically built into the evaluation design. The goals also constitute the frame of reference in which the behavioral scientist evaluator must assess the degree of relevance of the program methods.

Since so much of the evaluation is dependent on a clarification of the goals of the program, it would appear to be incumbent upon the behavioral scientist to at least raise the question as to how well the goals are related to the kind of problem which the program is attempting to change. Evaluating a program thus begins with identifying and analyzing its goals.

What is Included and What is Excluded From Evaluation

The one question which most clearly

distinguishes the behavioral scientist role from the data analyst role in evaluation is the degree to which the evaluator participates in the decision as to what, if anything, should be excluded from the evaluation.

For a variety of reasons having to do with the original purposes of the evaluation and with the personality and attitudes of the sponsoring persons, certain aspects of the program or its outcome may not be considered appropriate objects of evaluation. The data analyst may be able to accept such limitations uncritically by applying his technical skill to evaluating only those parts of the program which have been approved for study, and either to draw limited conclusions or let someone else write the conclusions.

The scientist on the other hand will want to establish the right to include as objects of evaluation those aspects of the program which, in his professional judgement, are necessary to accomplish a meaningful evaluation. He may feel that he can conduct such an evaluation by including certain minimum aspects without necessarily including all relevant aspects. Most evaluation is no doubt partial in this sense. It is, for example, not uncommon for evaluation studies to exclude from their assessment such variables as the personalities and competence of staff and still contribute to an understanding of the program's effectiveness.

But the crucial question must be faced by the potential scientist evaluator: does the exclusion of certain aspects of the program or its outcome as proper objects of study, render a meaningful result highly unlikely? If so, then the behavioral scientist will find an appropriate way of limiting or ending his role in the proposed evaluation.

Aspects of a program which might be excluded, in spite of the efforts of the behavioral scientist to include them, are: comparison of explicit and implicit goals; a detailed description of what the actual program consists of; the relevance of methods to goals; criteria of success most directly related to goals; a system of classifying staff which is relevant to their attitudes and behavior in carrying out their activities; a system of classifying clients relevant to the problem which the program is aimed at meeting; and the relevance of the goals to the nature of the problem and its cultural context. It is likely that many if not most of these program variables are given only minimal attention in the typical approach to evaluating programs for addicts.

In a given study, the exclusion of any of these variables could jeopardize its potential value to a greater or lesser degree. Such contextual constraints on the scope and nature of the evaluation to be permitted are encountered in almost every field. Because of the peculiar historical and political context in which addiction programs have been carried out in this country in the last several decades, the seriousness of these constraints are especially visible and can serve to alert evaluators in other fields to their possible detrimental influence.

What Should Be The Criteria of Success

Because of these constraints, traditional efforts at evaluating the effectiveness of treatment programs for addicts in this country have largely consisted of follow-up studies of addicts released from a hospital or other agency to determine the percent who were still off drugs. While these studies have increasingly included non drug-use data on adjustment variables such as employment and arrests, the drug use variable has almost universally been given preeminence.

In view of the routinely high rate of failure found when the primary criterion of success is drug abstinence, other kinds of secondary criteria have occasionally been used. These include the number of contacts with the treatment agency, completion of treatment, length of time off drugs before relapse, and proportion of time off drugs out of the total time since discharge.

It was not until 1964, to our knowledge, that an evaluation of an addiction program explicitly presented social functioning variables as alternatives to the primary success criteria of abstinence. In that study, conducted by one of the authors of this paper, success was defined in terms of increased conventional behavior and reduced criminal behavior. Drug use was recorded as a possible explanatory variable but was viewed as conceptually independent of success.[3]

This approach, which we have referred to as a rehabilitation model, is now being used by a number of programs with built-in evaluation. The criteria of success, however, vary with respect to the old bugaboo -- abstinence from drugs. Most of the current programs include abstinence from drugs other than methadone as a significant criterion of success.

One exception to this use of a new abstinence criterion is the Narcotic Addiction Demonstration Center of Southwestern Fairfield County in Stamford, Conn. Use of additional drugs is viewed there primarily as a matter of general research interest and clinical concern and only secondarily as a criterion of success.

The more common pattern of using abstinence from all drugs but methadone as a criterion of success equal in importance to those of "working" and "staying out of trouble with the law" is understandable when we take into account one fact: that powerful political pressures impinge on most addiction treatment agencies which have experimented with methadone maintenance. While the evaluator may or may not be able to resist these pressures, we submit that it is his obligation nonetheless to raise questions about alternative goals, methods, and criteria of success.

How Should Success Criteria Be Measured

Once the criteria of success have been conceptualized, the evaluator must decide what observable behavior should be used to indicate or measure the degree to which each success criterion has been achieved.

If, for example, being employed has been established as one criterion of success, how shall this be measured? One might use the client's own report as to whether he is employed or not. Self reported behavior by addicts, however, is often rejected as a reliable measure of success. Thus, one might scrutinize pay checks or obtain verbal or written confirmation by employers as such a measure.

Similarly, criminal behavior is often measured by arrest records. The unapprehended crime, however, may go unmeasured if the addict is not queried about it. A methodological question underlying this dilemma is whether one can get more valuable data by establishing the kind of relationship with clients which may potentially produce full and accurate information or by relying on "objective" records of client behavior which are somewhat less than comprehensive.

Moreover, problems of professional ethics should be considered in deciding on whether to use subjective reports, objective records, or some combination of the two. The implications for the field of addiction of choosing one set of measures rather than another should be carefully thought through. For evaluation is a form of action-research and its design and implementation reflect assumptions and values regarding the nature of the problem and its treatment, whether the evaluator is aware of these reflections or not.

A most cogent and controversial illustration of these methodological and ethical problems consists of the use of urine testing to measure the success criterion of abstinence from drugs. This method, called thin layer chromotography, is typically used in programs which include as one of their goals, either abstinence from all drugs or abstinence from all drugs except legally administered methadone.

Various reasons may be cited for the use of urine testing. These include the reputed tendency of addicts to prevaricate, the reputed readiness of addicts to belittle any program which does not use urine testing, and the need for some form of coercion or rational authority in treating addicts.

Of course, the last reason is an open statement of a treatment philosophy which incorporates thin layer chromotography as a necessary treatment modality. It should be considered, in such cases, as a method of treatment to be evaluated and not simply as a method of evaluation.

When urinalysis is not directly part of the treatment program, its pros and cons as a measure of abstinence can be appropriately weighed. Its advantages are that it doesn't require development of an open relationship between client and evaluator, it may meet the expectations of most of the clients, and, at least in the short run, it may well provide more reliable information about use of most types of proscribed drugs.

The disadvantages of urinalysis are that it fails to provide information about cocaine and alcohol (which the users of these tests claim they are also concerned about) and it can, under certain circumstances, be manipulated. But most important, we submit, it constitutes a procedure in which the client participates in a process of self-degradation.

Some might take issue with this judgement by pointing to the daily weighing of overweight clients and to the periodic urine testing of diabetics. The crucial difference, as we see it, resides in the criminal stigma which adheres to addiction but not to obesity and diabetes. In the case of addiction, thin layer chromotography can truly be called "guilt by urination."

While each evaluator must make his own judgement about the ethics of this procedure, at the least the question should be raised. Eventually, of course, the effects on addict clients and on public attitudes of urine testing should be carefully assessed. In the meantime, it is sufficient to point out that there are powerful political pressures for including urinalysis in addiction treatment program evaluations. This is perhaps the most telling argument in favor of its use. Nonetheless it is significant to note that the one program we know of which is evaluating a methadone maintenance program without thin layer chromotography reveals a higher rate of illegal cocaine use as measured by self-reported drug use than does a major program which depends on urinalysis which does not test for cocaine.

How Should Recipients Be Classified

The aim of evaluation, according to one research evaluator, is as follows: "As diagnostic classifications and treatment goals and methods are more sharply defined . . . the focus of the evaluative question is likely to be sharpened so that we may no longer be asking, how effective is psychotherapy . . ." -- (or, we might add here, any kind of intervention) -- ". . . but rather, how effective is such-and-such kind of treatment in producing such-and-such changes in such-and-such kinds of people."[4]

An evaluator is generally alert to the possibility that the rate of program success, whether high or low, may be to a significant degree the result of a skewed sample. The difficulty arises in determining the most appropriate systems for classifying the clients. Clients may, of course, include such diverse targets as persons with a problem, the general public, agency personnel, decision makers, etc.

In the addiction field in this country,

programs for the most part have been aimed at addicts. The classification systems used have typically been of personality variables of these recipients of program efforts. These classes are generally presented as predisposing factors in addiction.

If one assumes that treatment of the predisposing conditions in the host is an effective approach to curing a chronic problem, then psychiatric classifications offer one logical approach to characterizing clients. However, that assumption does not always hold. In the field of medicine, for example, progress in controlling disease is sometimes achieved without a clear understanding of its etiology.

Moreover, addiction is a complex sociomedical problem which involves deep social as well as psychological roots and resultants. We believe it is primarily because cultural constraints have dictated for so long that the de facto goal of treatment be complete cure -- that is, abstinence -- that classification systems have almost exclusively focussed on psychological variables. While this is one legitimate approach, it tends to preclude innovative approaches to treatment and alternative goals in place of abstinence.

Once a rehabilitation model is used to establish improved functioning as a goal of treatment, then the shortcomings of an exclusively psychiatric approach are apparent. When high priority is assigned to the goal of improved social functioning (for example, increased involvement in conventional areas of living and decreased involvement in criminality), then relevant new social classifications are essential.

Such an approach has been followed in two evaluation studies in which one of the authors has served as director. A classification of addict life style adaptation was developed based on two independent social characteristics -- the degree of an individual's involvement in the conventional world and in the criminal world.

A typology was established consisting of four distinguishable life styles: "Hustlers" -- those who conform to the stereotype of the addict as high in criminality and low in conventionality; "Conformists" who are low in criminality and high in conventionality; and two mixed groups -- "Two-Worlders" who are high in both criminality and conventionality, and the "Uninvolved" who score low on both of these dimensions.

The value of such classification is twofold. First it assigns addicts to groups which present very different types of problems for the community and for the treatment agency. Second, the typology provides a key set of criteria for measuring success in terms of improved functioning and more positive adaptation by addicted persons. Since improved physical and mental health are also goals of a comprehensive rehabilitation program, there is need for additional classifications of addicts in terms of their patterns of drug use and their psychological adjustment.

Our point, however, is not that the above classification of life-style adaptation should be universally used in evaluating addicts. More effective classifying schemes are sure to emerge. The important point is that the choice of an exclusively psychological approach to classifying addicts tends to place the evaluator in the role of a passive data analyst unless he is fully aware of the influence of the ambivalent cultural context on the goals, methods, and success criteria which form the basis of his evaluation and unless he is prepared to raise relevant questions about alternatives. One group of evaluators of California's civil commitment program did just that in stating that: ". . . commitment programs for addicts can be considered at this time an interim procedure between a totally punitive and evolving non-punitive approaches to the issues of drug dependence, though perhaps they will persist as an alternative for those who are not helped by other programs. Implicit in this view is the expectation that alternative approaches will be explored and encouraged."[5]

As we suggested earlier, addicts are only one target of intervention for programs of addiction control. Other targets, perhaps of equal or greater importance at certain points in time, are professional attitudes, agency postures, and public policies themselves. Measuring the success of such efforts is far removed from testing the urine of individual addicts. How we define the problem to be attacked is the ultimate yardstick which we must apply in evaluating programs aimed at the problem.

Other Contextual Considerations

Other questions involving the contexts in which treatment and evaluation are carried out should be raised by a conscientious evaluator. Time allows but brief mention of two of the more important inquiries. Who should do the evaluation is a question of obvious importance. Our own answer is that it is more important that the evaluator assume the role of behavioral scientist than whether the evaluator is an "insider" or an "outsider." We have seen situations in which both inside and outside evaluators of addiction programs have served solely as data analysts and other situations in which each have served as behavioral scientists.

Another source of difficulty might be avoided if the evaluator ascertains ahead of time how the findings of the evaluation will likely be disseminated. While acceptable reasons may exist for limiting the scope and nature of the distribution of results, implicit reasons which appear unjustified to the evaluator should be elicited and challenged.

The purposes of evaluation should generally include the feedback of findings -- both positive and negative -- to the program personnel. Here the evaluator-scientist role is most clearly that of an action-researcher. This phase of evaluation calls for interpretation by the evaluator and may culminate in specific recommendations for modifying the program's goals, methods or procedures.

Summary

In this paper we have tried essentially to make three interrelated points about evaluation. First, we submit that the case of narcotic addiction points up the degree to which the social and cultural context, -- in this case, in terms of punitive public policies and ambivalent community attitudes, -- can narrow the range of permissible goals and methods of treatment and the purposes and design of evaluation as well.

Given the criminal status assigned to narcotic drug use by our society, it is logical to aim programs at stopping the criminal behavior of drug use and to utilize involuntary methods of treatment and of measuring success. When addiction programs are, to a greater or less degree, shown to fail, the availability of psychiatric classifications of clients, legitimate as they might be, can too easily suggest that inadequate personalities are the major factor in this failure and thus divert attention from basic contextual sources of failure external to the individual.

This leads to our second point. Evaluators in designing their studies of programs in addiction, and no doubt in other fields, are inevitably and deeply involved in the action arena. Their evaluation designs and the criteria and measures they select will tend to reinforce the basic assumption of the addict as a criminal (regardless of any mental health facade) or, conversely, to challenge that assumption. These action consequences of seemingly objective research decisions can not be easily avoided. They are likely to characterize program evaluation in other fields as well.

Our third point, then, is that it is a basic obligation of the evaluator to consider the impact of the community context in which programs operate, to challenge basic assumptions which may have no foundation in fact, and to raise questions about alternative formulations of the problem and alternative approaches to controlling it. The evaluator, in short, if he is indeed to evaluate meaningfully, must assume -- regardless of his professional discipline and training -- the role of behavioral scientist.

REFERENCES

1. See, for example, Bigman, Stanley, "Evaluating the Effectiveness of Religious Programs", Paper read at Meeting of Religious Research Fellowship in New York City, February 14, 1958.

2. See Brotman, R., Meyer, A., and Freedman, A., "An Approach to Treating Narcotic Addicts Based on a Community Mental Health Diagnosis," Comprehensive Psychiatry, Vol. 6, No. 2, (April) 1965, pp. 104-118.

3. See Meyer, A. and Brotman, R., Some Problems and Prospects in Evaluation of Patient Care for Narcotic Addicts, paper presented at the annual retreat of the New York Medical College Research Society at Kiamesha Lake, N.Y. on November 13, 1964.

4. Herzog, Elizabeth, Some Guidelines for Evaluative Research. (Washington, D.C.: U. S. Children's Bureau Publication No. 375), 1959, p. 80.

5. Kramer, John, Bass, Richard, and Barecochea, John, Civil Commitment for Addicts: The California Program, (Corona: California Rehabilitation Center), 1967, p. 10.

EVALUATION OF NATIONAL FAMILY PLANNING PROGRAMS THROUGH GROUP-SPECIFIC FERTILITY PATTERNS

Samuel M. Wishik, International Institute Human Reproduction, Columbia University

As the title of this paper indicates, it focuses on evaluation of large scale Family Planning (FP) programs, usually encompassing an entire country or a major portion of a country. The components of evaluation of a national FP program might be represented by the following complex: I(KAP/F/S). The KAP constitutes the immediate concentration of the program's effort to inform people (that is improve their Knowledge), change their Attitudes and value systems (such as about family size), and increase the Practice of contraception among them. These are immediate objectives which are only incidental to what might be called the intermediate goal, that of lowering Fertility. But neither is lowering of fertility an end in itself. It is a step toward obtaining the benefits that would ultimately accrue to Society and to its members and families. It is obvious, however, that observed changes in any of these parameters could be accidental or a factor of time alone and may not be attributed to the program unless cogent and significant correlations with program Input can be demonstrated.

The present paper focuses on the F or fertility portion of the complex. A previous paper[1] reported on methods of drawing from administrative and service data estimates on the amount of contraceptive practice in a population. In process are studies on methods of utilizing data on prevalence of contraceptive practice to derive numbers of births averted thereby, in accordance with the statement $P_eF \rightarrow B_I$. The amount of contraceptive practice (P) of a measured degree of use-effectiveness (e) in a population of women of a certain level of fertility expectation (F) in the absence of contraception (or of the added contraception contributed by the program) will point toward estimating the number of births averted by the input of that program (B_I). Because of the many unknowns in the statement, we are not yet ready to use that approach. Instead, the present paper utilizes a different tack and tries to measure changes in fertility behavior more directly. The aim is to develop an instrument that can give prompt indication of fertility change and that can be applied to a population as a whole as well as to large sub-groups and therefore can be used as an index for evaluating the effectiveness of different amounts and types of program input over periods of time.

This is merely another effort in addition to the many ways in which demographers have attempted and have in part succeeded in producing more useful indicators of fertility change than the crude birth rate. Desirable criteria of such an instrument include promptness, simplicity, feasibility and reliability, as well as validity. The possible value of the method here proposed rests in part on the concept and in part on the method of obtaining the data. In a well developed country that boasts of reasonably complete coverage of vital statistics, available data can be subjected to the methods of tabulation and analysis here presented. In the absence of such vital statistics, the relatively simple single retrospective interview method here described would be resorted to.

One aspect of the method is by no means a new idea, that is to group women by certain common characteristics and see whether women with those characteristics show a reduction of fertility over periods of time. Among theoretically relevant factors that might be considered, such as education, income and urban living, two only were chosen. These are the woman's age and the number of living children she has. It is reasonable to believe that these are two factors that influence a couple's actions which do or do not lead to another pregnancy, as well as being related to the physiologic fertility potential of the family. In addition, the number of living children at the time of a retrospective interview seems to be infomation that could be obtained with greater accuracy than other types of pregnancy history, such as total number of pregnancies, total number of births including fetal deaths, or total live-born babies no longer surviving at the time of the interview. With the substitution of number of living children for parity, this parameter was labeled "Parenthood Status" and the groups of women were classified according to both their age and parenthood status (A/PS).

Although separated into sub-groups according to A/PS classification, the method aims at assessing the fertility patterns in a total population group for comparison over time intervals. The population under comparison might be those dwelling in a selected geographic area, such as a district of a country. Or, they might be sub-groups identified for some reason that might have bearing on the question under study, such as ethnic or religious differences. The comparison of changes in fertility patterns from time one to time two would have meaning if the groups being compared with each other during this time interval are assumed or have been found to be reasonably similar at the outset in such relevant considerations as contraceptive knowledge, attitudes and practice, as well as family size ideals or aspirations. If the objectives of the program focus on certain women such as those of high parity or those who have recently given birth to a baby, groups of such women would be compared with similar groups over the time period. But in no sense is it intended that this method be applied to differentiate changes in fertility behavior among women who are contraceptive accepters, adopters, users or practicers, whatever term is used, as compared with those who are not accepters, adopters, users or practicers. It would be fallacious to compare those who practice contraception with those who do not, on the assumption that any differences in fertility behavior found after a suitable time period could be attributed merely to that contraceptive practice. What is the nature of the fallacy?

In simplified terms, one might categorize couples in three groups.

a) The couple want pregnancy now, whether on cultural or psychological basis, conscious or less fully recognized, articulated or less clearly planned. The net effect is that they want or need a pregnancy at this time.

b) The couple is indifferent, apathetic or fatalistic. The net effect is a lack of apparent, real or, one might say, appropriate concern about the question.

c) The couple definitely does not want a pregnancy. They want not to have a pregnancy. In the absence of available family planning service or advice or just education about contraception, what would be the fertility behavior of this third group as compared with the first or even the second?

There are conflicting considerations. On the one hand, it may seem to be a reasonable hypothesis that the degree of objection to pregnancy, in this instance most marked in group c, is somewhat correlated with the previously demonstrated fecundability. If there were two women, each married ten years and without access to contraception, the one who has three children is more likely to have a fourth than the other who has only one child is likely to have a second. At least the former may be more concerned about the possibility of another pregnancy at that time; she may feel "more vulnerable". Whatever the differential physiologic fecundability, it is also a reasonable hypothesis that the fertility performance level will be lowered in some subtle way, despite the absence of a FP program, in proportion to the shift to the right on this a-b-c scale. The term used here, "in the absence of a FP program", means the situation as it existed without that planned program effort, whether the situation was one of practically complete absence of modern contraception or whether some part of the population was using traditional or even newer contraceptive methods. Some individuals in almost any population group will have adopted contraception, usually more so among the better educated and upper socio-economic groups. In other words, dichotomizing users from non-users prejudices fertility expectations, though we do not know how much or for certain in which direction.

The fallacy of making fertility comparisons between users and non-users in response to specific program input is similar to the situation in which the United States tobacco industry has with such apparent effectiveness argued against evidence about the damaging effects of cigarette smoking. They contend that men who are prone to cancer have the personality whereby they prefer to smoke cigarettes. This type of argument is even more appropriate to the contraception-fertility question, since contraception is directly aimed at affecting fertility, whereas the cigarette smoker is not usually consciously trying to shorten his life with lung cancer or coronary thrombosis or cerebrovascular accident. The tobacco question is partially satisfied by such consistent quantitative correlations between the amount of smoking and the incidence of certain diseases that it is difficult to conjecture that there exists a subtle gradation of personality as expressed by the amount of dependence on cigarettes rather than that the smoking behavior is the significant contributor to the pathological outcome.

It does not seem reasonable that one can overcome the argument, however, with respect to contraception and fertility as easily by showing differential fertility behavior among contraceptive practicers and non-practicers unless one instituted the controlled experimental design whereby all the women were offered contraception, but, among those who came up to the very threshold with evidence of desire to go through with it, only some are given the help and others are turned away. It is hardly likely that this type of research is feasible. The closest to this and the one that is considered in the present presentation is a matching of geographic areas or populations as mentioned before and the introduction of different amounts of family planning program efforts, whether contrasting some with none or more with less.

It is inevitable in the early stages of the development of a large national FP program that some "natural experiments" of this kind occur. It is not likely that a program can be launched simultaneously in all areas or developed universally at the same pace. Some areas will get service before others or more than others as well as different types of program effort. Unfortunately, the factors that usually contribute to such differential program investment are also relevant to fertility potential or behavior as well as the population's readiness to accept family planning. For example, programs tend to move first into the more accessible areas among the readier population groups.

On the other hand, a "natural experiment" may occur for reasons of staff differences rather than area differences, One district may be luckier in the quality, dedication or energy of its FP officer and staff. Accidental occurrences of illness, recruitment or staff turnover may produce differences. Searching for considerations of this kind constituted one basis for the selection of some of the areas in which the Survey of Fertility Patterns was done in Pakistan. Because this was a first run (after pretest) of an instrument not yet validated, the initial survey was considered in the nature of research on survey methodology rather than having as its primary purpose the obtaining of definitive national data. Therefore, no attempt was made to do a scientific sampling of the country as a whole. Instead, the respondents were selected about equally from East and West Pakistan. Each of these two provinces was further subdivided into northern, middle and southern areas, corresponding roughly to major ethnic and sometimes language differences as well as gross agricultural and industrial characteristics of these large regions. Approximately one-third of the respondents in each wing of the country were selected from the two largest cities, Karachi in West Pakistan and Dacca in East Pakistan. In the remaining two-thirds, the respondents were rural residents. 18 interviewers and 6 supervisors completed interviews of 12,339 women in eight weeks.

The selection also aimed at seeking for evidence of differential fertility response to differential reports of program accomplishments. In Pakistan, each district (there are more than 50 such districts in the country) submits a monthly report concerning the amount of contraceptive supplies distributed, the number of intrauterine devices inserted, the number of packets of oral contraceptive pills distributed and the number of surgical sterilizations performed. Each district compiles its data from reports it receives from the FP officers at the next lower geographic level, who in turn receive reports from their local area subordinates. In each of eight such predominantly rural districts, four local areas were selected on the basis of their being near the lower or upper end of the range of reported program accomplishments. High and low achievement areas were chosen in each district, on the assumption that the reports reflected in general terms the amount of uptake of contraception among the populations in those respective areas. The aim was to see whether differential fertility could be detected by the method when extremes of fertility might be anticipated. This and other calculations will be the subject of other reports.

In the geographic zones selected, area sampling of households was done. In each household, women between 15 and 50 years of age and who had been married and living with a husband at any time during the previous 5-year period were interviewed. The usual interview methods for collecting household membership information were followed. Copies of the household and individual respondent questionnaires are attached to this report. Standard practices were followed in selection, recruitment, supervision and monitoring of interviewers.

The interview with each respondent was quite brief, 15 to 20 minutes. The interviewer averaged 16 completed interviews per day. Four pieces of information were obtained.

1. The woman's age at the time of the interview. All the tricks in the trade were called upon to obtain this difficult bit of information with as much reliability as possible, especially by relating the woman's life experience to major political or catastrophic events with which everyone in the community was familiar.

2. The number of living children, that is the number of children born to the respondent who were still alive at the time of the interview. Effort was made to include children living away from home and, for obvious reasons and for others that will become clearer as the discussion of this paper proceeds, special attention was given to be sure not to miss infants under one year of age.

3. The number of children born alive during the most recent five-year period and the year in which each one of those was born. Here, special effort was made to fit the time of birth of each child into the twelve-month periods into which the 5-year study span was divided. Each study year ran from July 1 to June 30. Efforts were made to fix time of birth and age of child with the help of discussion about seasons, harvests and religious holidays as well as important events.

4. The number of deaths of children during the most recent five-year period and the year in which each death took place. Information was also obtained as to which one of the children died. For example, was it the infant born the same year, or a child born more than five years ago? Except for being interested in the total number of living children at the time of the interview, regardless of their ages, no attempt was made to obtain detailed data on children over five years of age or on the woman's pregnancies further back than five years.

It is of interest to interpolate here that in the first round of this survey, information about birth and death of children was separated into half-year periods so that there were ten such periods during the five-year span of the retrospective interview. When the data were analyzed, it was found, as had been reported earlier in Pakistan and elsewhere, that there was such a definite seasonal fluctuation in the timing of the birth of babies that it became difficult to delineate half-year trends. Therefore, half-years were merged and the data were analyzed in terms of full years. In future survey rounds, the practice of using one-year periods will be followed.

The interview findings were coded in a simple row of five boxes, each representing a one-year period, working backwards from the cut-off date of June 30, 1967. (Table 1) The occurrence of a birth or death in any one of the five boxes was noted. The woman was given her A/PS classification at the time of the interview. From this, the woman's age in each of the preceding years was noted, simply by deducting one year for each. The PS designation in each of the five years was then noted by assigning to each of those years the number of living children the woman had at the <u>beginning</u> of each period. A woman who reported four living children at the time of the interview and who had an infant born in the most recent year would have had a PS designation of three at the beginning of that year. If a death had occurred in any year, the PS was increased retroactively by one since that child had been alive at the beginning of that period. If both a death and a birth had occurred in a given year, that year's PS would be the same as that of the subsequent year, since the birth and the death neutralized each other and the number of living children had not changed from one year to the next. Consequently, the code now confers upon each respondent an A/PS classification for each of the five years in which she was an eligible candidate.

In effect, the method compares women with themselves as an overall group from one time period to another. These same women demonstrated certain fertility behavior in previous time periods and now within the group, the question is raised, "Are they behaving differently from the way they did before"? As will be described, this is in a sense a cohort and yet it is not. The same group of women are assessed over a range of years, but the women in each time period are grouped according to what their age and parenthood status were at that time period. One group of women who were 35 in the most recent year, would not be compared with themselves at 30, but with those women who had been 35 five years earlier. If there has been any change over the five years, the change has occurred among a single total universe of women without extracting from that universe those who did or did not respond by contraceptive practice to the education or service offered by the national FP program.

<u>Method of Tabulation</u>

For each of the five years, a grid was set up to distribute the women according to their classification in that year by age and parenthood status. Although the information was obtained in single units for number of living children and for year of age, the women were grouped in order to reduce the number of cells in the grid (Table 2). Of course, the type of grouping selected depends upon the particular population and the distribution of the data in that population. In the United States, a family of two children is quite different from one of three. In Pakistan, on the other hand, it was found convenient to set up the following PS groupings: no children, one child, two or three children, four or five children, more than five. For age, the traditional five-year groupings were utilized, although in further analysis of the data it may be decided to combine groups still further. It is possible that the women may be grouped as young, middle child-bearing years and old, with the cutoffs below 25 and above 34. It may also be found desirable to separate the youngest group under 20 and those 40 and over as being atypical and usually having small numbers of cases. Further study will determine this in accordance with the general objective of bringing the method down to the crudest level that will give useful information on trends. Obviously, the fewer cells in the grid the smaller the size of the sample called for in the survey. Ten thousand respondents was found quite adequate in the Pakistan survey. How much smaller than this the sample can go depends on the more particular types of rates sought in each instance.

For each of the five years then, there now exists distribution into the different cells of the grid of all the respondents who were of child-bearing age in that year. In each cell, the number of women constitutes the denominator. The births that occurred to those women in the same year constitute the numerator of a rate, the rate of the number of births that occurred in that year to women of a given PS classification in that same year. It now becomes possible to make comparisons of group-specific fertility behavior over the five years. For example, if among 100 women age 25 to 29, with two living children at the beginning of the study year 1962 (July 1, 1962 to June 30, 1963), 30 babies were born that same year, how does this compare with the number of babies born to 100 women who were that age and who had the same number of living children during the year 1967 (July 1, 1966 to June 30, 1967)? Trends of group-specific fertility performance can be studied.

It is clear that the measurement of fertility during a one-year period is an arbitrary limitation of time. The present method also permits answering the question, "Among women of a given PS classification, what is their fertility performance over a longer period of time than one year, for example, two, three, four or five years"? Table 3 indicates the years of the retrospect for which one-year to five-year fertility rates can be calculated. Obviously, such calculation of fertility rates for progressively longer periods increases accordingly the number of women in each cell who have had more than one birth encompassed in the fertility rate assigned to the total group, although of course, this occasionally occurred in the one-year calculation. Consequently, highly fertile women contribute disproportionately to the total. The utility of the different rates will be ascertained with further experience.

Another interesting by-product of the method may turn out to be an extremely important one. This is the attempt to use the data to relate survival experience of infants with intervals between the time of their births and that of an older sibling. The study gives indication of which child died, such as an infant who is born and dies in the same year, or a child who dies in one year and had been born in one of the other previous years of the five-year retrospect, or a child who dies during the five-year period. With careful attention to comparable periods of observation prior to and after birth, mortality rates of babies will also be derived according to age of older sibling in relation to time elapsing before the next younger baby was born. What does close spacing mean to the baby who is followed and to the one who follows? Obviously, other factors that are correlated with short interval must be kept in mind. It is not implied here that interval alone is the basis for mortality differences that may be found.

We should here like to discuss some possible artifacts that may exist in the method and hope that it will be possible to find ways of testing for them.

1. It is well known by census takers that women respondents often forget to tell about the youngest baby. In the Pakistan surveys, the interviewers were cautioned especially about this. They asked repeatedly about whether there is a little baby in the home. The baby seems less likely to be omitted in a culture where breast feeding is almost universal. More often than not, the baby is with the mother and can be seen by the interviewer even if the mother forgets the baby sleeping on her back. The possible error of forgetting the youngest baby could be a very serious one because the very heart of the purpose of the method is to identify promptly a lowering of recent fertility in a population subjected to a new large national effort. This question will receive attention and possible check in the second and subsequent rounds of the survey, as will be described below.

2. What possible artifacts might be induced by the error of ages of babies being thrown into adjacent years from the ones in which they were actually born? In general, one might say that this should balance itself out, because the three-year old infants would be thrown into either the two-year group or the four-year group and the four- and the two-year groups would in turn throw a similar number back into the three. A mathematical question arises as to whether such neutralizing exchange is as probable in the first and last years of the five-year retrospect as it is in the three middle years. The first year follows an indeterminate period whereas the most recent year is juxtaposed to the future and has room only for backward exchange. In consideration of this possible mathematical artifact, it is important that the interviews occur close to the cutoff date. If there is an artifact for this reason, one would guess that, if anything, it would favor a higher birth rate being recorded in the most recent year. Again, this possible artifact will be subjected to check at the second round of the survey.

3. The usual danger exists of falling into round numbers. The care with which the interviews were conducted in the Pakistan study, especially the utilization of outstanding public events in trying to determine age, may have helped to minimize peaking at round numbers. There is also an impression, which needs validating, that in a culture that is in part numerically illiterate or where numbers are less profoundly a constant feature of everyday living, falling into round numbers is less of a research hazard than in the more number-oriented societies. In trying to answer this question, it is important to differentiate the extent to which the professional or interviewing personnel introduce the tendency toward round numbers in their approach rather than this being the product of the respondents alone.

4. When seeking a very prompt indicator, it would be better to relate fertility patterns to estimated time of conception than to birth. Consideration is being given to what would be required to do this.

5. Study is also in process of the inferences that can be drawn from the use of different types of time bases in the comparisons, such as: annual trends, earliest versus most recent year, most recent versus average of previous blocks of years, blocks before and after initiation of a national program and use of overlapping periods of different lengths. Consideration is also being given to the

appropriateness of producing various types of overlapping combinations from data obtained from different rounds of the survey.

6. The most obvious question in the use of a retrospective method is, "What is the error of recall and does it increase with the length of the retrospective period"? One can hypothesize that the error for the first year which was five years before the time of the interview is greater than that for the most recent year and that there might possibly be a steady gradient of such diminishing error as recency is attained. One might also conjecture whether there is differential error for birth as compared with death and for birth of surviving versus non-surviving infants. It is more than likely, however, that the net effect of error of recall would tend to exaggerate recent fertility rather than the opposite. Therefore, any reduction in fertility that seems to be manifest in the data derived from the method might reasonably be considered as minimized, in this respect at least.

It is obvious from the nature of the method being described that its value would be enhanced by periodic repetition, in order to extend the observation of trends. A second round of the survey would also have the research function of answering questions about the instrument. In Pakistan, this will take place early in 1969, approximately 18 months after the first one. An annual periodicity had been projected but did not prove possible. Regardless of the periodicity, the second round will be so designed that comparable data on coterminous periods will be obtained. A five-year retrospect will be used as in the first round, but the period will be brought up to date so that approximately the first year will be dropped and a more recent year added. No attempt will be made to revisit the same respondents or households as of the first round. Each of the years covered will have been one year further back in retrospect in the second round than it was in the first. With fully random samples in successive rounds, it is therefore theoretically possible to compare findings for a given year when data for that year represent varying lengths of recall. This is a rather unique opportunity to try to answer this old and thorny question.

In an attempt to simplify the method still further and in accordance with recent experiences of other workers, particular attention will be given to shorter recall periods. How much can be learned and inferred from the simple questions, "Was there a birth in the most recent year"? and "How long a time has elapsed since the last birth"? Further studies over the years with attempts at checking by repeated rounds might give understanding of the length of recall period that is optimum in different settings as well as the extent of increasing error with its increasing length. It is our impression that with the arbitrary five-year cutoff, the first or oldest year in the retrospect is of itself a kind of "buffer" between the previous marriage period not analyzed in detail and the retrospective period included in the tabulation. If this proves to be the case, it may be necessary to have a five-year retrospect in the interview but utilize the data only of the most recent four years. Such use of a buffer might also extend into future decisions to shorten the total retrospect below five years.

Although future survey rounds will not seek a strict cohort, it may well be desirable to go back into some of the same communities. Whatever is done along these lines will be primarily for methodology research and supplementary to the clear intent to use completely random samples in the second round. Many aspects of the benefits of the repetition of the method and comparison of data among different rounds will come up for question in the future. How much variation is introduced by two rounds that cover a given year as compared with four or five rounds for that year? Is there differential reliability between types of respondents, urban vs rural? What are the distribution patterns of the ages of the women or of their PS as well as A/PS classification? Therefore, would the denominators in the different cells differ considerably before one looks at variations in the numerator fertility data.

The measurement of group-specific fertility rates gives only indirect clues to the occurrence of increased spacing or prolongation of intervals between pregnancies. For this purpose, the data can also be analyzed according to pregnancy-free intervals, whether closed by a subsequent pregnancy during the five-year period of retrospect or whether the intervals remain open. A variety of analyses are under way and will be the subject of a separate report.

It has been suggested that the data can be subjected to simpler numerator analysis as described by Ravenholt and Fredriksen.[2] Consideration will be given to the types of supporting data needed to draw inferences from the group-specific fertility numerators alone.

Without waiting for the checks that will be obtained from future rounds, the method lends itself, in part at least, to other types of reliability assessment. How does the retrospective interview method compare with concurrent interviewing over time where this has been done and with the recording of vital data when the recording is adequate and reasonably reliable? Even in places where vital statistics are admittedly incomplete and inadequate, it would be of interest to find out whether the data and the inferences being drawn from them are comparable to those obtained by retrospective studies in such countries. The method is now being subjected to a number of such comparisons. Only one will be mentioned here.

In Pakistan during four of the five years covered by the present report, a project has been going on called P.G.E. or Population Growth Estimate. This consisted of a representative sample of approximately 26 areas of the country in which were placed resident registrars who made monthly or quarterly reports of vital events, births and deaths in the households in their sample. The data from that project, which you might say were obtained prospectively although they were intermittently retrospective through frequent home visiting, are being subjected to tabulations according to the same method for comparison. The author would welcome communications and possibly collaboration with investigators who may wish to replicate the method in different settings. A manual of instructions on the Fertility Patterns Survey method is in production.

Although the present paper aims at presenting the method as a research problem, it is of interest to make a brief reference to certain findings of the Pakistan Survey. A drop in fertility seems to have occurred in the most recent year (July 1, 1966 - June 30, 1967). The intensified national Family Planning Scheme had been launched in September 1965. The most striking reduction in reported fertility, perhaps in the order of almost 20%, occurred among older women, especially in mid-thirties and later, who had more than three living children. This is not an unexpected first response to a new effort.

SUMMARY

A method is presented for retrospectively obtaining information on fertility performance during a five-year period and treating the data as if the women were being observed prospectively in groups classified according to their age and number of living children. The method was used in a survey of over 12,000 women in Pakistan. A second round survey will be done to test the method's reliability. Improved methods of data collection on fertility are needed to help in the prompt assessment of the effectiveness of national family planning programs.

Bibliography

1. Wishik, S. M. "Indexes for Measurement of Amount of Contraceptive Practice", (In Press).

2. Ravenholt, R. T. and Fredriksen, H., "Numerator Analysis of Fertility Patterns" Public Health Reports 83:6 (June 1968), 449

Table 1 Example of Coding

	July 1, 1962 – July 1, 1963	July 1, 1963 – July 1, 1964	July 1, 1964 – July 1, 1965	July 1, 1965 – July 1, 1966	July 1, 1966 – July 1, 1967	July 1, 1967
	1st year	2nd year	3rd year	4th year	5th year	
	birth	birth death		death	birth	
A/PS at beginning of each year	28/3	29/4	30/4	31/4	32/3	Interview A/PS 33 years 4 living children

Table 2 A/PS Distribution Grid for each of the Five Years

		Number of living children-Parenthood Status (PS)				
		0	1	2-3	4-5	over 5
age groups (A)	19 & under					
	20-24					
	25-29					
	30-34					
	35-39					
	40 & over					

Table 3 Group-specific Fertility rates in each cell of grid can be derived for the following:

1-year fertility rate for groups as of beginning of years 1,2,3,4,5
2-year fertility rate for groups as of beginning of years 1,2,3,4,
3-year fertility rate for groups as of beginning of years 1,2,3
4-year fertility rate for groups as of beginning of years 1,2
5-year fertility rate for groups as of beginning of year 1

Fertility pattern
Study Questionnaire:
Part I.

WP/FP/APS/H/8/67

HOUSEHOLD INFORMATION

Village/Household/Respondent No.
(Supervisor to fill)

1. Date of interview: Month ____________ Day ____________

2. Name of village: ____________

3. (a) Name of head of household: ____________

   (b) Caste of head of household: ____________

   (c) Name of father of head of household: ____________

4. Address: ____________

5. How many people slept in this household last night? ____________

6. How many of them are women who are now married? ____________

   Names:

   (a) ____________ Wife of ____________

   (b) ____________ Wife of ____________

   (c) ____________ Wife of ____________

   (d) ____________ Wife of ____________

   (e) ____________ Wife of ____________

==

(For interviewer only)

7. How many women were finally interviewed from amongst the total married women listed in Question 6? ____________

8. If any of them were not interviewed, give reasons for each: ____________

Remarks:

Fertility Pattern
Study Questionnaire:
Part II.

INDIVIDUAL RESPONDENT INFORMATION

WP/FP/APS/V/8/67

Village/Household/Client#

1 2 3 4 5 6 7

Do not mark in code boxes

9a. What is your name? ______

b. What is your husband's name? ______

10a. How old are you? ______

b. How old were you at Partition? ______

c. [Interviewer's opinion about age :] ______

11a. How long ago were you first married? ______

b. [If less than 5 years, nearer to :] ______

12. How many living children do you have? ______
[Do not include children born since June 30, 1967]

List the names of all of these children, starting with the youngest.

(a) ______

(b) ______

(c) ______

(d) ______

(e) ______

(f) ______

(g) ______

(h) ______

(i) ______

(j) ______

8 ☐ ☐ 9
10 ☐ ☐ 11
12 ☐ ☐ 13
14 ☐ ☐ 15

13. Mark the ages of all living children under 5 years, starting with the youngest.

4 Years		3 Years		2 Years		1 Year.		Under 1 Year.	
Nearer		Nearer		Nearer		Nearer		Nearer	
5 Years	4 Years	4 Years	3 Years	3 Years	2 Years	2 Years	1 Year	6 Mths. or more	Under 6 mths.

(Section below for coding only Note: Code numbers in boxes.)

Deaths	17	18	19	20	21	22	23	24	25	26	
All Births	27	28	29	30	31	32	33	34	35	36	[# in 14,15]
# living children at beginning of period	37	38	39	40	41	42	43	44	45	46	○
Age of mother	47-48	49-50	51-52	53-54	55-56	57-58	59-60	61-62	63-64	65-66	○
Eligibility	67	68	69	70	71	72	73	74	75	76	[# in 8, 9]

16 ☐

14. Have you had any children who were born alive to you but later died?

(even if they only lived for a few hours).

No ______ Yes ______ How many ______

a. Most recent death: How long ago? [Nearer to] ______

If less than 5 years ago; how old was child at time of death? ______ [Nearer to] ______

b. Next most recent death in past 5 years:

How long ago? ______ [Nearer to] ______

If less than 5 years, how old was child at time of death? ______ [Nearer to] ______

c. Next previous death in past 5 years:

How long ago? ______ [Nearer to] ______

If less than 5 years ago, how old was child at time of death? ______ [Nearer to] ______

[For interviewer to answer]

15. In your opinion, how reliable is the information given by this woman? [Mark with a tick]

1) Very reliable ______

2) Reasonably reliable ______

77 ☐ 3) Not reliable ______

16. Remarks:

Print name of interviewer: ______

Signature: ______

Signature of Supervisor ______

DISCUSSION

Ira H. Cisin, The George Washington University

On the surface, the Kelman-Elinson paper is a rather humble statement. We are presented with a picture of highly skilled and sophisticated research workers who seem to be reminding us that evaluation is the toughest kind of research, that it is even tougher than usual in their situation, and that they are working hard to overcome a long catalog of special difficulties. The tone is almost plaintive; on the surface they seem to be asking for sympathy. Well, there's nothing wrong with a good sad story -- with a note of the old college try, and an optimistic tone. If they wanted sympathy, they could certainly have mine.

But I suspect that behind this humble facade lurk some tough-minded men who have come here to teach us an important lesson. Behind the plea for sympathy Kelman and Elinson are issuing a challenge to the entire field of evaluation research; they are telling us not only about their inadequacies but also, and more importantly, about ours. In short, I see an important didactic message in this paper; these authors are trying to teach us something, if we will only listen.

Without putting words in the authors' mouths (they don't have to subscribe to my interpretation if they don't want to), I suspect that they are telling us that, behind the surface humility is an attack on the established ways of thinking about evaluation research; I suspect that they are telling us that our models, our paradigms, our smug and scientific dogma for evaluation research may simply not be applicable to many real life situations. It is not so much that their situation is difficult, perhaps uniquely difficult; rather I suspect they are saying that their situation may be more typical of the real world than the sterile easy situations for which our models are appropriate. I suspect they are saying that we are being forced back to the drawing board to cope with the fact that we can deal with textbook problems, but that the book solutions don't fit the real world.

For example, we tend to think of the evaluator as standing off from the action. He is the objective outsider, disinterested, with nothing at stake in the evaluation process. He studies the system but he is not part of it. Kelman and Elinson remind us rather forcefully that, in real life systems, the evaluation team is likely to be a part of the total system. How often is an objective evaluator called in from outside? And how often is his objectivity protected from the influence of his participation, his prospects of future consultation, the political and social pressures of the system. The Kelman-Elinson model reminds us that evaluation exists as a subsystem of the total system, integral to the operating system, embedded in it, subject to the joys and sorrows of the operating system, while the evaluator is trying to educate the other members of the system to the necessity for objectivity in evaluation, he is subject to suspicion, fears, threats, and pressures. In other words, the image of the evaluator as an uninvolved and temporarily employed expert, a non-participant temporary observer of the operating system is unrealistic. Kelman and Elinson are reminding us that not only don't we have solutions for this problem; we have refused to face it as a problem.

Derivative from our notion of the evaluator as external to the operating system is our concept of evaluation as a kind of grading system. We move in, remain uninvolved, and when we're through, we award a grade. We are testers, not teachers. Kelman and Elinson remind us that, in reality, the evaluator can rarely hit and run in this way. Not only does he award the grades after test but, in their situation (and I suspect in a great many others) the tester is also the teacher. In the kind of long-range program that Kelman and Elinson are dealing with, the evaluation is a continuous process, the evaluator's input is continuous and presumably timely. Thus, the evaluator's role is one of diagnostician in a system under continuous flux. I share the authors' dislike of the term inspector; but inevitably the evaluation sub-system will be feeding news of success and failure into the operating system -- and would any of us be so pure as to say that the operating system ought not to respond to this news?

We are fond of talking about the need for clear definition of objectives; Kelman and Elinson remind us that while our notions may be appropriate to single-stimulus one-shot programs, the evaluator of the long range program will, of necessity, have wished upon him the objective of continual diagnosis and correction, contributing to the modification of the operating program during the test period. Simple models of the evaluation process simply do not cope with this reality.

Quite apart from the problem of the intermediate impact of the evaluation, we have another kind of problem in the design of evaluation. Traditional evaluation has its basic origins in the art of experimental design and it should not surprise us that we have not solved in evaluation a problem that has never been solved in experimental design: the problem of value for effort. We don't really need to do something-vs.-nothing experiments: we know by now that with a few exceptions the fertilized plot yields more corn than the unfertilized plot, that trained soldiers hit more targets than untrained soldiers. The something-vs.-nothing experiment makes sense only if we seriously entertain a hypothesis that the experimental stimulus will accomplish nothing or may actually do harm. In the ideal situation of course, we know the price of the fertilizer and we know how much we can get for our corn so we can stage optimization trials to find the most profitable level of fertilization. But if the criterion cannot be measured in the same metrics as the experimental input, optimization trials cannot be conducted.

Now Kelman and Elinson remind us that a great deal of evaluation is analagous to the something-vs.-nothing experiment; or more precisely to a more fertilizer vs. less fertilizer experiment, where the less-fertilizer stimulation is known to be inadequate. We want to improve a service program, so we pour money and staff into it. Then we ask to what extent we have achieved our objective. Now since the objectives we're interested in can almost never be expressed in money terms, the outcome of the evaluation is almost inevitably that the new more expensive program is better than the old less-expensive program. That phrase "additional funds" is like your grandmother's chicken soup: it couldn't hurt. Now this is the situation Kelman and Elinson describe: the program is hyped up with additional staff and additional money; the new program is extremely likely to be better than the old program. But, to the more important question of optimization; to the more important question of value; to the more important questions: are we getting our money's worth, are we spending the additional money in the most intelligent way or even in some reasonably intelligent way? -- to such questions our models of evaluation simply offer no guidance.

Finally I want to compliment these authors for bucking a trend. In our quest for easily measurable criteria in evaluation, we tend to duck some of the difficult problems we would face if we took our criteria seriously. In medical and quasi-medical programs, we have devoted a great deal of effort to the refinement of measures of effectiveness of treatment. There is a real danger that we will lose sight of the fact that the success of all service programs depends in the long run on the satisfaction of the persons served. There is a market research aspect to the rendering of service. The expectations, desires and indeed the satisfaction of the recipient cannot be ignored. I think Kelman and Elinson are due special applause for their broad view of the criterion problem, with the medical expert-opinion of the treatment on the one hand, and the patient's non-expert, uninformed, but overwhelmingly important opinion on the other.

Kelman and Elinson have tackled an extremely tough job; they have embarassed us by pointing out how little of our pious traditional thinking about evaluation is applicable to their problem; they have the courage of pioneers; and I wish them luck.

DISCUSSION

Charles R. Wright, National Science Foundation

I am indebted to Alan Meyer and Stanley Bigman for a most instructive and thoughtful presentation of problems faced in the evaluation of narcotic addiction control programs. What I am about to say is in no way intended to detract from the value of their case but rather to raise some questions about further generalization that their paper suggested to me.

From a methodological point of view I wonder what, if anything, is unique about evaluations of narcotic addiction control programs. Are there methodological problems which are specific to the evaluation of such programs or, if not specific, are at least more serious, troublesome, or pronounced there than in other kinds of evaluation research? Are there special practical problems in the execution, administration, and official or public acceptance of evaluation research in these programs that are not found in others? This line of analysis, if followed, would enhance our appreciation of the distinctiveness of evaluation efforts in this problem area. But better still, once certain features of evaluation research in this field are identified and examined they may highlight certain formal methodological problems that can be found in similar if not identical shape in evaluation research in other fields. Some illustrations follow.

A common problem faced in evaluation research is specification of the temporal aspects of effects. The program is to achieve certain effects within some period of time. Now consider, for the sake of an example, that one of the goals of a narcotic addiction control program is to get addicts to abstain from the use of narcotics, or detoxification. For how long must a former addict abstain from the use of narcotics before he constitutes a successful case in the evaluation of the program -- a few months, a few years, forever? Here we have an unusual set of circumstances by which each failure, that is, each case of return to the use of drugs, becomes readily apparent as a black-mark against the program while each success remains dubious, a potential failure until the end of the temporal cut-off point (by extremely harsh criteria, until the day he dies!). By contrast, evaluations of many other types of campaigns, e.g. getting out the vote, disbursing information, or getting people immunized, work with relatively fixed time limits within which the program's effects are contained. Each success within the time period is clear -- the man votes, he improves his knowledge about the topic, he comes to a clinic for a vaccination. If there is any uncertainty, it lies on the side of undetected cases of success that appear, because of imperfect information, to be failures. Is this problem of an open-ended time dimension unique to narcotics program evaluation or are there similar instances in other areas of evaluation research? If there are, how have they been solved and what are the methodological implications?

A second common problem in the conceptualization of effects during evaluation research, as described by Hyman and others (1), is the need to specify the locus of effects in terms of regions and sub-regions. By regions we mean whether the goal of the program is to affect the individual, an aggregate of individuals, a group, a total community or society, or some combination of these. Sub-regions might specify, for example, whether the effects are in terms of overt conduct, values, opinions, attitudes, motivation, interests, information, or some other phenomena. Insofar as narcotic addiction control programs have as one of their goals the detoxification of the individual, then the locus of effects is individual conduct. But it is a very special kind of conduct, ceasing one kind of illegal behavior. Evaluation research in this instance faces serious ethical and methodological problems in attempting to obtain data about the individual which make him liable to legal punishment. Again this problem may be encountered in a class of evaluation studies, e.g. delinquency programs or criminal rehabilitation, in contrast with evaluations of such positively sanctioned activities as adult education, citizenship training, public information campaigns, and the like. I will not go into the matter of ethics. Considering only one methodological issue, we are faced here with the need to insure what Suchman (2) calls "subject validity." How can the researcher minimize the respondent's understandable reluctance to disclose information that he regards as harmful, even legally incriminating, to himself? Comparisons of various attempts to achieve high validity in this field and in others would be instructive.

One more example of problematic dependent variables: the authors suggest that a new rehabilitation-based model of narcotic control programs stresses "improved social and psychological functioning and improved physical health within the limits of chronic disability (i.e. without regard to whether the patient or client is on or off drugs)." From a research viewpoint I find this goal as unclear operationally as are many others. It would be useful to hear just how such goals are to be translated into concrete measures that permit evaluation of successes and failures. Also it would be instructive to discuss the methodological implications of this new goal. It seems to imply the need for at least a before-and-after experimental design, probably also a longitudinal one involving continuing measurements. Question: just how much change in social and psychological functioning and physical health, separately or in combination, is necessary to qualify as significant improvement? And for how long must such change persist?

In conclusion I have two comments which time does not permit me to develop here. First, it seems imperative that we learn more about the factors that affect the utilization of evaluation research findings, whether in narcotics addiction control programs or others. Second, I agree wholeheartedly with the authors that the evaluator must be free to act as a professional scientist rather than as a technician (or data analyst) under the control of others, if evaluation research is to fulfill its promise of useful, scientific contributions to the rational understanding and planning of social action programs.

REFERENCES

1. Hyman, Herbert H., and Charles R. Wright, Evaluating Social Action Programs, in P. F. Lazarsfeld, W. H. Sewell, and H. L. Wilensky, eds., The Uses of Sociology, New York, Basic Books, 1967.

2. Suchman, Edward A., Evaluative Research, New York, Russell Sage Foundation, 1967.

DISCUSSION

Ray H. Elling, University of Connecticut

The paper presents an attempt carried out in another culture to develop a methodology for recording number of births in a family. The investigators are to be commended for the innovative approaches and diligence of the effort as covered in the paper. The long-range goal of developing valid measures of the effectiveness of family planning programs is an obvious priority in today's world.

There are several observations to make about the effort as reported in the paper.

First, it is nearly amazing to find this number of interviews completed on the average per interviewer per day. Like the man who lost his watch someplace else but continues to look for it under the street light, we should seriously consider moving our expensive field surveys to Pakistan -- regardless of the problem.

The investigators present their approach and instrument as a success in general. It should be noted that they have not tried other approaches (or do not report them) to lend support to the comparative efficiency and effectiveness of this approach and instrument. We really know nothing of its validity except in a "face" sense, though this seems good.

There is probably more to be done in this approach to recall of past events along lines of the focussed interview.[1] The investigators do employ memory tripping events important in the local culture and this seems desirable. If the culture were intimately understood, more could be done along these lines. This leads me to a final point.

It is high time that the potential contributions of social scientists to cross-cultural research and programs were taken seriously. It is surprizing that this important endeavor apparently did not engage these talents in an integral, continuing fashion. A quotation from Foster seems appropriate to conclude this point and these remarks:

"At the risk of appearing to end on a negative note, I feel compelled to point out that almost all the problems stated in this paper and the conclusions drawn result from ex post facto accounts of what went wrong. In most cases the best we can do as social scientists is to suggest reasons for these failures; to argue that if different approaches had been taken, the outcome might have been different; and to hope that someday we can cooperate in the planning of programs from the beginning and follow them through to completion. The critical need in intercultural health programs is for social scientists to work in the field, to observe, experiment, compare, and test. The precise form of their work is secondary in importance to their active participation in programs. Failures to date would appear to be about equally attributable to the unwillingness of administrators of health programs to entrust social scientists with significant responsibilities, and the unwillingness of social scientists to commit themselves for long periods of study, often under conditions of personal hardship (especially for their families), in types of work that have less status for them and their colleagues than traditional research. Yet, until these interdisciplinary barriers are greatly lowered, we shall make only halting progress in evolving working relationships between medicine and social science."[2]

REFERENCES

[1] R.K. Merton and P. Kendall, "The Focused Interviewer", Amer. Jrl. of Sociol., 51, 541-42.

[2] George M. Foster, Problems in Intercultural Health Programs, N.Y.: Social Science Research Council, Pamphlet 12, April 1958, P. 45.

XI

CONTRIBUTED PAPERS I

Chairman, GEORGE F. KEARNS, U. S. Bureau of the Census

THE QUALITY OF REPORTING SOCIAL SECURITY NUMBERS IN TWO SURVEYS

Mitsuo Ono, George F. Patterson, and Murray S. Weitzman*
U.S. Bureau of the Census

Introduction

This analysis presents findings on the rate of reporting and the extent of "correct" reporting of Social Security numbers (hereafter designated as SSN's in this paper) by persons included in a sample of households obtained in two mail-out/mail-back content pretest surveys. These pretest surveys were conducted by the Bureau of the Census in May 1966 and April-May 1967 in connection with the planning of the 1970 Census of Population and Housing. In May 1966, the Bureau of the Census conducted the First Content Pretest in two areas--St. Louis Park, Minnesota and Yonkers, New York. This pretest survey universe contained approximately 4,700 households. A simple random sample of approximately 660 households (or about 14 percent of the total number of households) was selected for this study on SSN's. These households included about 2,000 persons of all ages. In April and May 1967, the Bureau of the Census conducted another pretest survey (the Second Content Pretest) at Gretna, Louisiana. The pretest survey universe was approximately 4,200 households. A simple random sample of 765 households was selected for this analysis. These households contained about 2,500 persons of all ages.

In both of these pretest surveys, all persons in the surveyed households who had an SSN (or a Railroad Retirement number) were requested to write in their SSN in the pretest schedules. Whenever a member of the household had no SSN, he was requested to fill in the "none" circle located directly to the right of the space for the nine-digit SSN. Respondents were requested not to guess their numbers, but to consult records, such as Social Security account cards or Federal tax records, whenever necessary. Thus, the collection of SSN's presented somewhat different reporting problems as compared with other Census questionnaire items that could be estimated without the need to consult records.

In this study, data intended for statistical tabulations were transcribed from the schedules onto special forms, verified, and forwarded to the Social Security Administration for processing. Statistical tabulations on the correctness of the SSN information were obtained from the Social Security Administration. Selected data from these statistical tabulations are presented in the appendix tables.

Summary of Findings

Among all persons 14 years of age or over, approximately 76 percent reported an SSN in the two pretest surveys. Fourteen percent marked the "none" circle indicating that they did not have an SSN and about 10 percent did not answer the SSN question. Approximately 77 percent of all persons 14 years old or over answered the question "correctly." This "correct" rate consisted of 67 percent of cases reporting that they had an SSN and 10 percent of cases reporting that they did not have an SSN. In other words, among persons reporting that they had an SSN, i.e., 76 percent, approximately 87 percent reported their "correct" SSN. For persons who reported they had no SSN's, i.e., 14 percent, 73 percent reported "correctly," that is, no SSN's were located for such persons.

For some of the persons reporting an incorrect SSN, no SSN, or who did not answer the SSN question, it was possible to locate their "correct" SSN's through additional effort. This effort resulted in improving the rate of "correct" response from 77 percent to 88 percent.

---

*Comments represent views of the authors and not necessarily those of the Bureau of the Census.

Note: Findings from another statistical study on Social Security numbers can be found on pages 18 and 19 in the following unpublished paper: "Interacting Data Systems and the Measurement of Income Size Distributions," by Joseph Steinberg, Social Security Administration, presented at the Conference on Research in Income and Wealth, Philadelphia, Pennsylvania, March 24-25, 1967. Suggestions by Mr. Joseph Steinberg of the Social Security Administration are gratefully acknowledged.

Table A.--REPORTING OF SSN's BY PERSONS 14 YEARS OLD AND OVER IN THE FIRST AND SECOND CONTENT PRETEST SURVEYS

Response	Total	Answered question			"Incorrect" answer possibly "corrected" by search	Residual (col. 3 minus col. 5)
		"Correctly"	"Incorrectly"	Information not available		
	(1)	(2)	(3)	(4)	(5)	(6)
All	2,962 (100.0)	2,276 (76.8)	676 (22.8)	10 (0.3)	332 (11.2)	344 (11.6)
Reported having SSN	2,257 (76.2)	1,974 (66.6)	283 (9.6)	--	88 (3.0)	195 (6.6)
Reported not having SSN	413 (13.9)	302 (10.2)	101 (3.4)	10 (0.3)	101 (3.4)	--
Did not answer	292 (9.9)	--	292 (9.9)	--	143 (4.8)	149 (5.0)

Note: Percentages in parentheses.

Source: Appendix tables 1 and 2.

Frequency of Reporting SSN's

The usefulness of obtaining SSN information in survey work depends on the percent of the population reporting SSN's. As shown in table B, persons 25 to 64 years of age reported that they had SSN's more frequently than did persons in the under 25-year or over 65-year age groups. A comparison of reporting rates of persons 14 to 19 years of age with that of persons 20 years old and over included in the Second Content Pretest survey is presented in table C. About 40 percent of persons 14 to 19 years of age reported not having an SSN as compared with 12 percent of persons 20 years old and over.

Table B.--REPORTING OF SSN's BY PERSONS 14 YEARS OLD AND OVER BY AGE GROUPINGS IN THE FIRST AND SECOND CONTENT PRETEST SURVEYS

Response	Total	Age groups					
		14 - 24	25 - 44	45 - 54	55 - 64	65 - 71	72 and over
Total	2,962 (100.0)	789 (100.0)	921 (100.0)	493 (100.0)	398 (100.0)	192 (100.0)	169 (100.0)
Reported having SSN	2,257 (76.2)	492 (62.4)	774 (84.0)	402 (81.5)	323 (81.1)	144 (75.0)	122 (72.2)
Reported not having SSN	413 (13.9)	209 (26.5)	56 (6.1)	34 (6.9)	48 (12.1)	36 (18.7)	30 (17.8)
Did not answer	292 (9.9)	88 (11.2)	91 (9.9)	57 (11.6)	27 (6.8)	12 (6.2)	17 (10.1)

Source: Appendix table 2.

Table C.--REPORTING OF SSN's BY PERSONS 14 TO 19 YEARS OF AGE AND PERSONS 20 YEARS OLD AND OVER IN THE SECOND CONTENT PRETEST SURVEY ONLY

Response	Percent	
	14 to 19 years	20 years old and over
Total......................	100.0	100.0
Reported having SSN.............	46.4	76.7
Reported not having SSN.........	39.6	12.2
Did not answer..................	14.0	11.0

Source: Special tabulations not included in appendix tables.

Table D shows that male persons 14 years old and over report having SSN's more frequently than female persons.

Table D.--REPORTING OF SSN's BY MALE AND FEMALE PERSONS, 14 YEARS OLD AND OVER, IN THE FIRST AND SECOND CONTENT PRETEST SURVEYS

Response	Total	Male	Female
Total...........................	2,962 (100.0)	1,386 (100.0)	1,576 (100.0)
Reported having SSN..................	2,257 (76.2)	1,150 (83.0)	1,107 (70.2)
Reported not having SSN.............	413 (13.9)	96 (6.9)	317 (20.1)
Did not answer.......................	292 (9.9)	140 (10.1)	152 (9.6)

Source: Appendix table 3.

Information in table E obtained from the Second Content Pretest survey indicates that household heads report more frequently having SSN's than "other" household members. About 86 percent of household heads reported having SSN's as compared with 62 percent for other members of households.

Table E.--REPORTING OF SSN's BY HEADS OF HOUSEHOLDS AND OTHER MEMBERS OF HOUSEHOLDS, 14 YEARS OLD AND OVER, IN THE SECOND CONTENT PRETEST SURVEY ONLY

Response	Total	Heads of households	Other members of households
Total........................	1,655 (100.0)	723 (100.0)	932 (100.0)
Reported having SSN...............	1,194 (72.1)	620 (85.8)	574 (61.6)
Reported not having SSN...........	271 (16.4)	36 (5.0)	235 (25.2)
Did not answer....................	190 (11.5)	67 (9.3)	123 (13.2)

Source: Appendix table 4.

"Correct" Reporting of SSN's

Persons reporting "correctly" their SSN's comprise two subgroups: (1) persons not only reporting having SSN's but also reporting them "correctly" and (2) persons reporting "correctly" not having SSN's. The proportion of "correct" reporting can be computed using either a population base including all persons 14 years of age and over in the pretest surveys or only persons 14 years of age and over in the particular population group to be analyzed. The proportion of "correct" reporting using the base covering all persons 14 years and over is found in table A. The proportions of "correct" reporting using the base covering only persons in the specified group are presented in tables F, G, and H, below.

Table F shows the proportion of "correct" reporting of persons reporting having SSN's and persons reporting not having SSN's. The proportions of persons who reported "correctly" were about the same for the various characteristics shown below except for persons in the age category 72 years and older.

Table F.--CORRECT REPORTING OF SSN's BY PERSONS 14 YEARS OLD AND OVER WHO REPORTED HAVING AND NOT HAVING SSN's IN THE FIRST AND SECOND CONTENT PRETEST SURVEYS

Selected characteristics	Reported having SSN and not having SSN	"Correctly" reported	Percent "correct" reporting
	(1)	(2)	(3) (2)/(1) X 100
All 1/	2,670	2,276	85.2
By age 1/			
14 to 24 years	701	583	83.2
25 to 44 years	830	720	86.7
45 to 54 years	436	385	88.3
55 to 64 years	371	316	85.2
65 to 71 years	180	155	86.1
72 years and over	152	117	77.0
By sex 1/			
Male	1,246	1,074	86.2
Female	1,424	1,202	84.4
By relationship 2/			
Heads of households	656	562	85.7
Other members of households	809	682	84.3

1/ Results obtained from both pretest surveys.
2/ Results obtained from the Second Content Pretest survey only.

Source: Appendix tables 2, 3, and 4.

Table G demonstrates that persons willing to report SSN's tend to report their SSN's "correctly." The proportion of persons reporting correctly did not vary much by different age groups, except for persons in the age group 72 years old and over.

Table G.--REPORTING OF "CORRECT" SSN's BY PERSONS 14 YEARS OLD AND OVER WHO REPORTED SSN's IN THE FIRST AND SECOND CONTENT PRETEST SURVEYS

Selected characteristics	Reporting SSN	Reporting a "correct" SSN	Percent "correct" reporting
	(1)	(2)	(3) (2)/(1) X 100
All 1/	2,257	1,974	87.5
By age 1/			
14 to 24 years	492	436	88.6
25 to 44 years	774	679	87.7
45 to 54 years	402	358	89.1
55 to 64 years	323	280	86.7
65 to 71 years	144	124	86.1
72 years and over	122	97	79.5
By sex 1/			
Male	1,150	1,010	87.8
Female	1,107	964	87.1
By relationship 2/			
Heads of households	620	537	86.6
Other members of households	574	511	89.0

1/ Results obtained from both pretest surveys.
2/ Results obtained from the Second Content Pretest survey only.

Source: Appendix tables 2, 3, and 4.

Table H shows that among persons reporting not having SSN's, about 73 percent reported the question "correctly." Female persons and "other" members of the household who reported not having SSN's tend to report this item more correctly than male persons and household heads.

Table H.--REPORTING OF "CORRECT" SSN's BY PERSONS 14 YEARS OLD AND OVER WHO REPORTED NOT HAVING SSN's IN THE FIRST AND SECOND CONTENT PRETEST SURVEYS

Selected characteristics	Reported not having SSN	No SSN was found	Percent "correct" reporting
	(1)	(2)	(3) (2)/(1) X 100
All 1/	413	302	73.1
By age 1/			
14 to 24 years	209	147	70.3
25 to 44 years	56	41	73.2
45 to 54 years	34	27	79.4
55 to 64 years	48	36	75.0
65 to 71 years	36	31	86.1
72 years and over	30	20	66.7
By sex 1/			
Male	96	64	66.7
Female	317	238	75.1
By relationship 2/			
Heads of households	36	25	69.4
Other members of households	235	171	72.8

1/ Results obtained from both surveys.
2/ Results obtained from the Second Content Pretest survey only.

Source: Appendix tables 2, 3, and 4.

Information Obtained From Additional Effort

Data in table I show the extent of additional SSN's located for respondents who either reported an "incorrect" SSN, or reported no SSN, or who did not answer the question. Among 283 persons who had reported "incorrect" SSN's, possible "correct" SSN's were found for about a third of these persons. Among 413 persons who had reported that they had no SSN's, possible correct SSN's were designated for about 25 percent of these persons. Among 292 persons who did not answer the question, possible SSN's were found for about half of these persons. Approximately two-thirds of the number of persons in the age category 14 to 24 years who did not answer the question were found to have possible correct SSN's. Among persons either reporting "incorrectly" or not answering the question, possible SSN's were located more frequently among male persons than for female persons. However, possible SSN's were located at about the same rate for both household heads and other members of households who either answered this question improperly or omitted the item.

Table I.--POSSIBLE SSN's LOCATED FOR PERSONS REPORTING SSN "INCORRECTLY," NOT HAVING AN SSN, OR NOT ANSWERING SSN QUESTION IN THE FIRST AND SECOND CONTENT PRETEST SURVEYS

Response	Total 1/	By age						By sex		By relationship 2/	
		14 to 24	25 to 44	45 to 54	55 to 64	65 to 71	72 and over	Male	Female	Heads of households	Other members of households
A. Persons reporting "incorrect" SSN	283	56	95	44	43	20	25	140	143	83	63
B. "Incorrect" answers possibly corrected	88	17	30	16	16	4	5	52	36	25	18
C. (B/A) net percentage	31.1	30.4	31.6	36.4	37.2	20.0	20.0	37.1	25.2	30.1	28.6
D. Persons reporting not having SSN	413	209	56	34	48	36	30	96	317	36	235
E. Possible SSN located	101	56	12	7	12	5	9	29	72	11	64
F. (E/D) net percentage	24.5	26.8	21.4	20.6	25.0	13.9	30.0	30.2	22.7	30.6	27.2
G. Persons not answering questions	292	88	91	57	27	12	17	140	152	67	123
H. Possible SSN located	143	58	44	22	9	4	6	75	68	33	57
I. (H/G) net percentage	49.0	65.9	48.4	38.6	33.3	33.3	35.3	53.6	44.7	49.3	46.3
J. All persons 14 years and over	2,962	789	921	493	398	192	169	1,386	1,576	723	932
K. Possible SSN located from search (B + E + H)	332	131	86	45	37	13	20	156	176	69	139
L. (K/J) net percentage	11.2	16.6	9.3	9.1	9.3	6.8	11.8	11.3	11.2	9.5	14.9
M. "Correct" reporting rate without search 3/	76.8	73.9	78.2	78.1	79.4	80.7	69.2	77.5	76.3	77.7	73.2
N. "Correct" reporting rate with search (L + M)	88.0	90.5	87.5	87.2	88.7	87.5	81.0	88.8	87.5	87.2	88.1

1/ Results from First and Second Content Pretest surveys.
2/ Results from Second Content Pretest survey only.
3/ Lines C through J, col. 3 + col. 9 (table 2).
Line C + line G (table 3).
Line C + line G (table 4).

Overall, possible "correct" SSN's were located for 332 persons, or for approximately 11 percent of the 2,962 persons 14 years of age and over included in both pretest surveys. This additional effort resulted in increasing the accuracy of the cases sampled from 77 to 88 percent. The highest gain in the improvement in the accuracy of reporting occurred in the 14 to 24-year-old age group.

<u>Summary</u>

In general, it can be concluded from this analysis that persons who are willing to report SSN's tend to report them correctly. The "best quality" of reporting SSN's can be expected for male heads of households who are 20 years old or over.

APPENDIX

Table 1.—ANALYSIS OF SOCIAL SECURITY NUMBERS COLLECTED IN THE FIRST AND SECOND CONTENT PRETESTS, BY AGE

Results of analysis	Percent— First and Second Content Pretests				Percent—First Content Pretest				Percent—Second Content Pretest			
	Total	14 years old and over	Under 14 years	Unknown	Total	14 years old and over	Under 14 years	Unknown	Total	14 years old and over	Under 14 years	Unknown
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
A. Total persons in pretest sample...........	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
B. Reported an SSN in pretest..	50.9	76.2	2.4	3.2	53.8	81.3	1.9	—	48.5	72.1	2.8	7.2
C. Correct SSN...............	44.5	66.6	2.1	3.2	46.9	70.8	1.8	—	42.6	63.3	2.4	7.2
D. Incorrect SSN.............	6.4	9.6	0.3	—	6.9	10.5	0.2	—	5.9	8.8	0.4	—
E. Reported <u>not</u> having an SSN in pretest.................	33.3	13.9	85.7	0.8	30.7	10.9	91.1	1.4	35.3	16.4	82.0	—
F. Possible SSN located......	2.8	3.4	2.0	—	1.8	2.0	2.1	—	3.5	4.5	1.9	—
G. SSN not located...........	29.6	10.2	81.6	—	26.9	8.1	83.7	—	31.8	11.9	80.1	—
H. Insufficient information provided.................	0.9	0.3	2.1	0.8	2.0	0.8	5.3	1.4	—	—	—	—
I. Did not answer SSN item.....	15.9	9.9	11.9	96.0	15.5	7.8	7.0	98.6	16.2	11.5	15.2	92.8
J. Possible SSN located......	3.2	4.8	0.2	—	2.7	4.1	0.2	—	3.6	5.4	0.3	—
K. SSN not located...........	6.6	5.0	11.7	1.2	4.3	3.7	6.8	1.4	8.5	6.1	14.9	0.9
L. Insufficient information provided.................	6.0	(1/)	(1/)	94.9	8.4	(1/)	(1/)	97.2	4.1	(1/)	(1/)	91.9

1/ Excluded from total since data for line L were not available in the Second Content Pretest.

Table 2.—ANALYSIS OF SOCIAL SECURITY NUMBERS COLLECTED IN THE FIRST AND SECOND CONTENT PRETESTS, BY AGE
(Combined totals)

Age	Total persons in First and Second Content Pretests	Reported an SSN in pretest					Reported not having an SSN				Did not answer SSN item			
		Total	Correct SSN	Incorrect SSN			Total	Possible SSN located	SSN not located	Insufficient information provided	Total	Possible SSN located	SSN not located	Insufficient information provided
				Total	SSN located	SSN not located								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
A. Total..........	4,515	2,296	2,009	287	90	197	1,503	126	1,338	39	716	146	300	270
B. Under 14 years.......	1,270	31	27	4	2	2	1,088	25	1,036	27	151	3	148	—
C. 14 years and over....	2,962	2,257	1,974	283	88	195	413	101	302	10	292	143	149	—
D. 14 to 24 years.....	789	492	436	56	17	39	209	56	147	6	88	58	30	—
E. 25 to 44 years.....	921	774	679	95	30	65	56	12	41	3	91	44	47	—
F. 45 to 54 years.....	493	402	358	44	16	28	34	7	27	—	57	22	35	—
G. 55 to 61 years.....	300	239	210	29	11	18	39	11	28	—	22	7	15	—
H. 62 to 64 years.....	98	84	70	14	5	9	9	1	8	—	5	2	3	—
I. 65 to 71 years.....	192	144	124	20	4	16	36	5	31	—	12	4	8	—
J. 72 years and over..	169	122	97	25	5	20	30	9	20	1	17	6	11	—
K. Not reported.........	283	8	8	—	—	—	2	—	—	2	273	—	3	270
L. Total..........	100.0	50.9	44.5	6.4	2.0	4.4	33.3	2.8	29.6	0.9	15.9	3.2	6.6	6.0
M. Under 14 years.......	100.0	2.4	2.1	0.3	0.2	0.2	85.7	2.0	81.6	2.1	11.9	0.2	11.7	—
N. 14 years and over....	100.0	76.2	66.6	9.6	3.0	6.6	13.9	3.4	10.2	0.3	9.9	4.8	5.0	—
O. 14 to 24 years.....	100.0	62.4	55.3	7.1	2.2	4.9	26.5	7.1	18.6	0.8	11.2	7.4	3.8	—
P. 25 to 44 years.....	100.0	84.0	73.7	10.3	3.3	7.1	6.1	1.3	4.5	0.3	9.9	4.8	5.1	—
Q. 45 to 54 years.....	100.0	81.5	72.6	8.9	3.2	5.7	6.9	1.4	5.5	—	11.6	4.5	7.1	—
R. 55 to 61 years.....	100.0	79.7	70.0	9.7	3.7	6.0	13.0	3.7	9.3	—	7.3	2.3	5.0	—
S. 62 to 64 years.....	100.0	85.7	71.4	14.3	5.1	9.2	9.2	1.0	8.2	—	5.1	2.0	3.1	—
T. 65 to 71 years.....	100.0	75.0	64.6	10.4	2.1	8.3	18.7	2.6	16.1	—	6.2	2.1	4.2	—
U. 72 years and over..	100.0	72.2	57.4	14.8	3.0	11.8	17.8	5.3	11.8	0.6	10.1	3.6	6.5	—
V. Not reported.........	100.0	2.8	2.8	—	—	—	0.7	—	—	0.7	96.5	—	1.1	95.4

Table 3.—ANALYSIS OF SOCIAL SECURITY NUMBERS COLLECTED IN THE FIRST AND SECOND CONTENT PRETESTS FOR PERSONS 14 YEARS OLD AND OVER, BY SEX

Results of analysis	Number		Percent	
	Males 14 years old and over Total	Females 14 years old and over Total	Males 14 years old and over Total	Females 14 years old and over Total
	(1)	(2)	(3)	(4)
A. Total persons in First and Second Content Pretest samples	1,386	1,576	100.0	100.0
B. Reported an SSN in pretest	1,150	1,107	83.0	70.2
C. Correct SSN	1,010	964	72.9	61.2
D. Incorrect SSN	140	143	10.1	9.1
E. Reported not having an SSN in pretest	96	317	6.9	20.1
F. Possible SSN located	29	72	2.1	4.6
G. SSN not located	64	238	4.6	15.1
H. Insufficient information	3	7	0.2	0.4
I. Did not answer SSN item	140	152	10.1	9.6
J. Possible SSN located	75	68	5.4	4.3
K. SSN not located	65	84	4.7	5.3
L. Insufficient information	—	—	—	—

Table 4.—ANALYSIS OF SOCIAL SECURITY NUMBERS COLLECTED IN THE SECOND CONTENT PRETEST FOR SPECIFIED PERSONS 14 YEARS OLD AND OVER

Results of analysis	Number		Percent	
	Heads of households Total	All other household members Total	Heads of households Total	All other household members Total
	(1)	(2)	(3)	(4)
A. Total persons in Second Content Pretest	723	932	100.0	100.0
B. Reported an SSN in pretest	620	574	85.8	61.6
C. Correct SSN	537	511	74.3	54.8
D. Incorrect SSN	83	63	11.5	6.8
E. Reported not having an SSN in pretest	36	235	5.0	25.2
F. Possible SSN located	11	64	1.5	6.9
G. SSN not located	25	171	3.5	18.3
H. Insufficient information	—	—	—	—
I. Did not answer SSN item	67	123	9.3	13.2
J. Possible SSN located	33	57	4.6	6.1
K. SSN not located	34	66	4.7	7.1
L. Insufficient information	—	—	—	—

STATISTICAL APPROACHES TO ASSESSING QUALITY OF MEDICAL CARE

Isidore Altman *

To set the stage for this presentation and to explain its rather different character -- in the sense that it is different from most presentations at an A. S. A. meeting -- let me quote from my letter to our program chairman, Mr. Elijah White.

"Some years ago, I published an annotated bibliography on methodology in evaluating the quality of medical care, 1955-1961.(1) A supplement covering 1962 through 1967 (now 1968) is . . . nearing completion. If you . . . feel something of this sort -- a description of what people are doing in this field (no statistics, no equations) -- might be suitable, I would be glad to submit an abstract."

I thought it might be of a little interest to a group like this, the Social Statistics Section, to hear of attempts to measure in some statistical way an intangible concept like quality of care -- to measure something which cannot even be defined, except that we all probably have some notion about what it is; and good care is something we all devoutly wish for when we get sick. I shall not attempt to define it myself, except to quote from Dr. Donabedian: ". . . the definition of quality may be almost anything anyone wishes it to be although it is, ordinarily, a reflection of values and goals current in the medical care system and in the larger society of which it is a part." (2)

The basis for my remarks is a bibliography I have been compiling (or, more accurately, have had excellent young women assistants compiling) since 1955. We began our work in 1961 but made 1955 our starting point for search of the literature. The University of Pittsburgh Press has published a collection of abstracts covering the years 1955-1961, inclusive, and we hope to publish another collection next Spring for the years 1962 through 1968.

If you reflect for a moment, it may occur to you that every paper that has to do with treatment or handling of patients has some relationship to quality. Writers on the subject of quality have had to wrestle with the problem of scope. We tried, therefore, to set some rules or criteria for selection. In the preface to our first publication, we said this:

"There are a great many papers appearing in medical journals which make a contribution to medical care because they present new drugs, new instruments and new procedures that are proposed and evaluated in terms of assisting in the diagnosis, treatment, or alleviation of a specific condition. Such items were generally omitted from this bibliography. For example, a paper describing the value of a new antibiotic would be excluded, but a paper which contained a method for determining the extent to which patients were being treated with antibiotics unnecessarily would be included. We are concerned in this bibliography with methods which would have more general applicability than the effect upon one specific condition. /_ The next is important because here writers differ._/ Studies of efficiency, costs, or utilization of services have been excluded unless it was the intent of the author to demonstrate the bearing of these aspects on quality. While the focus was on personal health services, many psycho-social aspects, nursing procedures and mental health were omitted."

As you might suppose, there were and are many borderline situations and difficult decisions to be made. We can only try to be reasonably consistent. And when you are charging a fee for your collection of abstracts, the customer does not mind a few extra abstracts; we have never had any complaints on that score.

Now let me, as it were, dispose of two more kinds of papers, one of which might be called non-statistical and the other "general statistical." As to the first kind, there are many papers in the literature which discuss one or more of the problems involved in measuring quality, with no data being utilized, except perhaps casually. While we try to avoid these in our bibliographies, a number of them do get in -- because of the cogency of their arguments or because they describe new systems, such as progressive patient care. Some are valuable review papers which decidedly warrant inclusion. A more or less typical abstract of a non-statistical paper reads like this:

"Changes in the social order such as population mobility, shifts in age composition and disease prevalence, and commercialization of the professions are discussed in relation to their effects on patient care and the search for better patient care. The author sees the need for cooperation among members of the health professions in order to improve the quality of care and he sets forth possible approaches toward the

---

* Professor of Biostatistics, Graduate School of Public Health, University of Pittsburgh. Presented before the Social Statistics Section, Americal Statistical Association, August 21, 1968.

achievement of this cooperation." (3)

The second kind of paper, what I have termed the "general statistical", represents the usual or most common approach where the collection of data is involved. Such a paper might describe a trial or an experiment where observations are made on two groups -- one treated the other untreated; one subjected to a new pattern of care, the other not -- and then the proportion of recoveries, or whatever the appropriate measure might be, is calculated. The analysis is based on simple rates, computed for different age groups or the sexes, etc. Sometimes a test of significance for the difference between proportions is applied, which all too often serves chiefly to clutter up the tabulations. (I am jesting, but only in part.) One of the best studies, which went beyond simple rates, says this:

"The methods employed throughout this study for testing the statistical significance of the observations to rule out the possibility of chance happenings consisted of standard procedures. These included contingency X^2 corrected for continuity, t-test for difference between two means, analysis of variance and covariance, t-test for significance of regression and correlation coefficients, and non-parametric procedures such as the sign test. Results stated as statistically significant were such that the probability of their occurrence by chance alone was always 5% or less. In some cases the exact probability was stated." (4)

I am going to kill two birds with one stone, in a way, by supplying illustrations within the classification scheme we have been following. Most attempts to assay the quality of some aspect of medical care fall into one of two broad groups: (1) observation of performance, directly or indirectly, and the application of some scheme of judgment, or (2) a gauging of the results presumably achieved. (Our published bibliography has two other sections, one calling attention to published sets of established standards of performance and the other containing a few unclassifiable items.) Time permits only an example or two from each of 5 sub-heads under the two major headings.

Under the heading Elements of Performance we had as a first sub-head "Audit, review, and evaluation." It seems only right to begin with a study that is becoming a classic in this field - - the study of general practice in North Carolina by Dr. Osler Peterson and his associates, from which I have already quoted. (4) This major study consisted basically of actual observations of physicians by other physicians as they carried on their practices. Practice was broken down into components which were scored -- by observing judges, in effect. Then, as part of the study, the findings on quality of work were shown to be associated with certain elements of education and training and other factors; that is to say, certain characteristics of the physician were predictors of professional performance (within limits).

Our second sub-head under Elements of Performance was "Suitability of the modality of care." The word "modality" refers to such kinds of care as home care, nursing home care, intensive care, physical and occupational therapy, and social services.

Here the before-and-after kind of approach appears to have proven the most useful. The functions the patient is able to perform or the activities he is able to handle might be coded periodically and measures of functional improvement obtained in some way.

The third sub-head under Elements of Performance is "Screening and case-finding." Here we generally have in mind community efforts to improve quality through prevention and timely treatment. Statisticians who have involved themselves with health examinations of substantial numbers of people have investigated the occurrence of false-positive and false-negative findings, with a view to determining the sensitivity and specificity of the tests which are employed. (5) Sensitivity "is measured by the percentage of all diagnosed cases that are screened positive by a particular test." Specificity "is measured by the percentage of the persons without the disease that are screened negative by the test in question." One statistician and his colleagues found that "The 'cheaper' tests frequently emerged as the most expensive ones when viewed in terms of effectiveness in identifying disease." (6)

Our second major category was the Effects of Care. In its most elementary aspect, the question one asks is, Did the patient recover? If he did, then presumably he received good care. (As an aside, it is possible that he recovered in spite of poor care or no care.) On a higher level, the approach has been through the examination of mortality and morbidity statistics--often the comparison of rates after the institution of some program of treatment or intensive study with the rates that had prevailed previously. Our first sub-head under Effects of Care is therefore

"Mortality and morbidity rates".

One notable paper here was an evaluation of a measure (the perinatal mortality rate) which is commonly accepted as an index of quality of maternity care. If only one measure had to be used, this study showed, the perinatal mortality rate was the figure to be used. But the figure itself needed to be adjusted for differences in population composition, and the user was warned to be wary of over-simplification resulting from the use of a single index.[7]

Our second sub-head "Patient expectations and satisfactions" gets over into human behavior, emotions, psychology, and the like. If the patient is happy about his care and he thinks or believes that he was the recipient of quality medical care, then ipso facto his care has been good. This seems to be a generally accepted notion, as judged from the papers we have abstracted -- and I would not quarrel with it.

A good deal of attention is paid these days to satisfaction with nursing care in the hospital. The most intensive study of this kind, probably, went at it by correlating satisfaction with different specific elements of nursing. Over 8000 patient forms and over 9000 personnel forms (usable ones) were collected in the course of the study. Patients and personnel other than nurses, it turned out, did not see eye to eye with the nurses themselves on which items of care best make for a happy patient.[8] But I get the impression that patient satisfaction is a very difficult thing to measure, and that the investigators are struggling.

As I indicated earlier, our final section was "Other" -- actually called, General Approaches. Few studies were so unclassified, mostly discussions. But one study here tried to get a picture of general quality of medical care in the community by assessing the amount of unmet meed. This was done through a household survey which inquired about symptoms. As you can begin to gather, quality is a most elusive concept.

Finally, I must call attention to two writers in this field and their works, with apologies to the other fine students whom I have not singled out. The first of these I consider our leading thinker on the subject of quality of medical care, Dr. Avedis Donabedian, Professor of Medical Care Economics at the University of Michigan, whom I mentioned earlier. Two very thoughtful and definitive papers of his (from which I am constantly borrowing) have appeared, respectively, in the Milbank Memorial Fund Quarterly July, 1966, Part 2, and in Medical Care, May-June, 1968.[9]

The other writer, Dr. Mindel C. Sheps, Professor of Biostatistics, Columbia University, School of Public Health, published her paper, "Approaches to the Quality of Hospital Care" back in 1955 in Public Health Reports.[10] But no paper since has approached it for clarity of exposition on the methodology of the kinds of studies we have been discussing.

I might mention that the "biggest study of them all" is currently under way at Yale's School of Public Health. Top-flight physicians in the School of Medicine are being questioned in great depth on how a long list of specific diseases should be diagnosed and treated. Out of this will evolve, I am sure, much clarification of what good medical care should consist of.

As I read our abstract and quite a number of the papers themselves, I get un uneasy impression of a lack of attention to statistical refinements in the studies which have been made; randomness, bias, reliability, validity, etc., do not receive the attention they should. But the field is in an early stage and the methods are still crude and rough and ready. It is a complex but intriguing field. These are "constraints" which may or may not attract you.

References

(1) Anderson, Alice J., and Altman, Isidore. Methodology in Evaluating Quality of Medical Care: An Annotated Selected Bibliography, 1955-1961. Pittsburgh: University of Pittsburgh Press, 1962.

(2) Donabedian, Avedis. Evaluating the Quality of Medical Care. Milbank Memorial Fund Quarterly 24:166-206 (Part 2), July, 1966.

(3) Simmons, Leo W. Change in the Social Order as it Affects Health Services. Hospital Progress 45:81-86, June, 1964.

(4) Peterson, Osler L., Andrews, Leo P., Spain, Robert S., and Greenberg, Bernard G. An Analytical Study of North Carolina General Practice, 1953-54. J. Medical Education 31(12): 1-165 (Part 2), Dec. 1956.

(5) Kurlander, Arnold B., Hill, Elizabeth H., and Enterline, Philip E. An Evaluation of Some Commonly Used Screening Tests for Heart Disease and Hypertension. J. Chronic Diseases 2:427-439, Oct. 1955.

(6) Clark, Thomas W., Schor, Stanley S., Elsom, Katharine O., Hubbard, John P., and Elsom Kendall A. The Periodic Health Examination: Evaluation of Routine Tests and Procedures. Annals of Internal Medicine 54:1209-1222, June 1961.

(7) Pakter, Jean, Erhardt, Carl L., and Jacobziner, Harold. Perinatal Mortality Rates as an Aid in Assessing Maternity Care. Amer. J. Public Health 45:728-735, June 1955.

(8) Abdellah, Faye G., and Levine, Eugene. Effect of Nurse Staffing on Satisfactions with Nursing Care. Hospital Monograph Series No. 4 Chicago, American Hospital Association, 1958. 82 pp.

(9) Donabedian, Avedis. Promoting Quality through Evaluating the Process of Patient Care. Medical Care 3:181-202, May-June 1968.

(10) Sheps, Mindel C. Approaches to the Quality of Hospital Care. Public Health Reports 70:877-886, September 1955.

METHODOLOGY IN A STUDY OF THE IMPACT OF WATER FLUORIDATION ON DENTAL PRACTICE */

Monroe Lerner, The Johns Hopkins University;
Bruce L. Douglas, Donald A. Wallace, and
Sylvia B. Coppersmith, University of Illinois.

A study of the impact of water fluoridation on various social and economic aspects of dental practice has been carried on since 1963 at the University of Illinois College of Dentistry. The background and plan of the study have been described elsewhere [3]. Both prior to initiation of the study and, based on several special "feasibility" surveys during its course, several major decisions involving research design and methods were made. It seemed worthwhile, even though retrospectively, to reconstruct and articulate the bases for these decisions in three major areas - the appropriateness of alternative study settings, how best to maximize the degree of cooperation with the study by the dentists, and alternative data-gathering strategies - since this exercise may be useful to students doing further work in this field.

I. THE STUDY SETTING

One of the most difficult decisions facing the study staff at the very outset was to select the most appropriate setting for the study from among the various possible alternatives. The two major alternatives considered included: 1) a national sample of dentists and/or patients; and 2) sets of matched communities, i.e., communities with fluoridated and communities with fluoride-deficient water supplies. If the latter of these two alternatives should be selected as most appropriate for this study (as in fact it was), the subproblem which followed was: Should communities with artificially or naturally fluoridated water supplies, or both, be studied?

When the present study was being conceived, i.e., during the early 1960s, relatively few communities in this country were enjoying the benefits of either artificially or naturally fluoridated drinking water. Furthermore, those drinking artificially fluoridated water had not experienced this innovation long enough for the occurrence of major changes in dental practice which might result from this change, at least not on any sizable scale. It was felt by the study staff, therefore, that a national sample of dentists even at best might not have included a sufficient number who had been practicing in fluoridated communities for at least some years to permit a reasonable degree of confidence in the study findings. To have obtained a satisfactory result by this procedure would have required a great deal of effort and expense, and might not even have been possible at all. But perhaps most important, water supplies involve <u>entire communities</u>, and therefore fluoridation presumably affects dental practice on a <u>community-wide</u> basis. Also, it was believed that dentists could more readily be induced to cooperate with the study if approached on a community-wide basis through their local professional organizations. To consider dentists apart from the communities in which they practice thus seemed to be an unreasonable procedure, and therefore communities, rather than a national sample of individual dentists, became the preferred study setting. These communities were to be matched on all relevant characteristics except the one: fluoridation.

The next problem in design was whether, as the most appropriate study setting, to select communities with naturally fluoridated or those with artificially fluoridated water supplies. By the early 1960s the value of fluoridation in controlling dental decay had been clearly established, and some investigators were already in addition pointing to its possible impact on dental practice. Kirby [5], for example, in 1959 had written that:

> "Prevention of dental decay [by fluoridation] will be accompanied, of course, by a reduction in the need for restorative dentistry. The private practitioner, then, will be able to devote more attention to problems brought on by occlusal wear and periodontal disease." */

Nevertheless, relatively few communities had had artificial fluoridation for even as long as a decade and it seemed clear that, although Kirby's predictions might hold in the future, the changes that he envisioned were not likely to have occurred as yet, at least not on a large scale. As a matter of fact, Kirby's speculations were deliberately stated in the future tense. And while Kirby had written in 1959, the situation was not much different in the early 1960s. On this basis, the authors of the present report rejected artificially fluoridated communities as the preferred setting for the study and chose naturally fluoridated communities instead.

Naturally fluoridated communities offered the advantage that they had possessed this form of fluoridation for a long period of time; it could reasonably be expected, therefore, that their patterns of dental practice would have been established for a long period of time also. While this appears to be a decided advantage, a disadvantage of using naturally

---

*/ Presented at the Annual Meeting of the American Statistical Association, Social Statistics Section, Session on Contributed Papers. Pittsburgh, Pa., August 21, 1968.

*/ [5], p. 87.

fluoridated communities was that it cannot be stated, at least with firm assurance and on an a priori basis, that natural and artificial fluoridation were likely to have had identical impacts on dental practice. For one thing, natural fluoridation had been in existence for a long period of time while artificial fluoridation represented a sudden change. Nevertheless, in the absence of any evidence to the contrary, and because it was known that their impact on dental health was similar, the assumption was made that the long-run impacts on dental practice of these two quite divergent forms of fluoridation would ultimately converge, and that the one might reasonably be taken, therefore, to represent the other. Natural fluoridation thus became the study setting of choice.

A perhaps logically extraneous but in fact practically important aspect of this decision was the matter of convenience and expense. The study team was located in Chicago. Illinois and the neighboring states of Indiana and Ohio are relatively rich in naturally fluoridated communities. It would therefore be convenient, and presumably relatively economical as well, to conduct the study in these states. While ideally this should not influence the decision, it actually became an important consideration. Also important, however, was the possibly biasing effect of the impact on dental practice of the very introduction of fluoridation itself, often after a referendum and usually accompanied by much publicity. Might not this publicity have attracted a special type of dentist to the community? The answer to this question is not known. However, natural fluoridation had existed for long periods of time without any such publicity and was therefore free of this possible source of bias.

Seven pairs of communities were selected, matched on relevant demographic and socioeconomic characteristics and differing only in that one community within each pair had natural fluoridation while the other had fluoride-deficient drinking water. To reduce the possibility of geographic and/or interstate variation perhaps affecting the results, six of the seven matched pairs of communities were selected from within the same state and all from within the three states of Illinois, Indiana, and Ohio. No large cities were included, and each pair was selected to minimize ethnic diversity as a possible source of disturbance.

The seven pairs of communities used in the present study were matched on the basis of population size, age-and-sex composition, family income, level of educational attainment, and years of residence in the community of its native-born population. */ All of these data were obtained from the 1960 United States Census of Population. In addition, the number of dentists, both general practitioners and dental specialists, practising in each community, the ratio of dentists to population, age of dentists, and the schools from which they had graduated were also considered in the matching process, but these factors were not crucial in each decision to pair specific communities.

The matched pairs were: Aurora and Rockford, Ill.; Kewanee and Centralia, Ill.; Marion and Sandusky, Ohio; Joliet, Ill. and Mansfield, Ohio; Elwood and Connersville, Ind.; Huntington and Shelbyville, Ind.; and Frankfort and Crawfordsville, Ind. (The first-listed in each pair had naturally fluoridated water at optimal levels.) During the final phase of the study the dentists in Rockford, a fluoride-deficient community, refused to cooperate and the community had to be withdrawn from the study. No other single community was suitable as a replacement, and it was necessary to substitute two communities, Freeport and Kankakee, Ill., for Rockford.

Three possible sources of disturbance could not readily be eliminated from the research design. These were as follows:

1) In many American communities not all persons drink water obtained from one central source. Some communities have more than one central source and many families drink water from private wells; the fluoride content of their drinking water may therefore be quite variable. The extent to which this was the case in the present study is not known.
2) Persons drinking fluoride-deficient water may nevertheless obtain the benefits of fluorides from other sources, e.g., topical fluoride application, fluoride tooth paste, etc. Again, the extent of this in the present study is not known.
3) The American population is highly mobile. At the time of the study, a very large proportion of the population in both sets of communities had not resided in their present community of residence during the first ten years of their lives.

In addition to these, one other possible source of disturbance which deserves mention proves that social change will not stand still, not even for social research. Thus between the time that the research design was formulated and the actual field work, four of the seven cities designated as fluoride-deficient introduced controlled artificial fluoridation of their water supplies. However, on the as-

*/ The authors of the present report are indebted to the National Opinion Research Center of the University of Chicago for assistance in making the selections.

sumption that such recent fluoridation would not as yet have affected dental practice substantially, no attempt was made to compensate for this change, e.g., by replacing these communities with others. Rarely, if ever, is it possible in real life to set up an experimental situation involving human behavior, even on an _ex post facto_ basis, from which all possible sources of bias, or disturbance, are removed, and the present study was certainly no exception to this rule.

2. INDUCING DENTISTS TO COOPERATE

Dentists in private solo practice, like physicians and other professionals, are generally believed to be somewhat less than willing to cooperate in social and behavioral research studies, especially where both their treatment methods and various financial aspects of their office practice are to be used as research data. The reasons for this reluctance appear to be at least these: Dentists in general are not likely to be familiar with or oriented toward social science and/or social research; they often resent what they regard as unwarranted intrusion into their professional affairs; their patient-records are considered to be, by professional ethics and often by law, confidential; and finally, their time is extremely valuable. */ How, then can they best be induced to cooperate with a social-research study?

One item that proved to be absolutely essential to obtain cooperation by the dentists in the present study was endorsement by local and state dental societies. The best cooperation rate that could be obtained without it, in two feasibility surveys that were part of the present study, was only 50 percent. But with this endorsement, 43 of 44 dentists in one community, or nearly 98 percent, indicated their willingness to cooperate.**/ However, endorsement was apparently a necessary but not sufficient condition for cooperation; even with it, most of the dentists in Rockford, one of the larger communities in the study, refused to be involved. If the cooperation rate is computed by including, instead of Rockford, the two communities which replaced it, 87 percent of all dentists cooperated. If Rockford is counted, however, the rate was much lower.

Since the time of dentists is generally conceded to be very valuable, an offer of remuneration was made to one-half of the dentists asked to cooperate in one of the feasibility surveys. No such offer was made to the other one-half. This offer, unexpectedly, appeared to make absolutely no difference in their decision, since the cooperation rate was identical in both groups. Also, only four of the 35 dentists to whom remuneration was offered accepted, and one of these donated the money to charity. However, an offer of remuneration was apparently very helpful in inducing _dental assistants and receptionists_ within the dentist's office to cooperate with the research. In some instances these people manifested obvious resistance to providing the researcher with the required records, or explaining where they might be found, and this despite their employer's expressed willingness. The dentists themselves often appeared to be reluctant to pressure them into cooperating. However, _merely the offer_ of remuneration resulted in cooperation in all instances, and in no case was the offer accepted.

3. ALTERNATIVE DATA - GATHERING STRATEGIES

Another question investigated here was this: Could a sufficiently large response rate be obtained from dentists by a relatively economical mail survey, or would it be necessary to use the more expensive method of personal interviews supplemented by photostatting designated patient-records or abstracting them in the dentist's office? One of the surveys conducted in the initial phase of the present study provided fairly substantial evidence that mail surveys could provide at best only a relatively poor rate of response, especially on treatment data.

This survey was conducted during the summer of 1963, and was aimed at the dentists practising in the seven pairs of matched communities. The names and addresses of approximately 400 dentists in these communities, and some biographical data about them, were obtained from the most recent issue of the _American Dental Directory_. A letter was sent to each dentist listed and a questionnaire was enclosed. The questionnaire had been designed with the assistance of staff members of the Bureau of Economic Research and Statistics of the American Dental Association and the letter contained an endorsement of the study by that

*/ The reluctance of dentists in this country to participate in research studies is evident from their low response rate to a mailed questionnaire from their own professional organization, the American Dental Association. Thus the _1965 Survey of Dental Practice_ reported a response rate of only 20.4 percent, _even though no treatment data had been requested_. [2]

**/ This extraordinarily high rate of willingness to cooperate was probably influenced by the fact that the community in question was a university town, so that the tolerance level for social research may have been greater among the dentists residing there than is usually the case elsewhere.

organization.

The first part of the questionnaire, contained on both sides of a single page, requested data from each dentist on biography, educational and professional background, and various social and economic aspects of his practice, including some financial items. These were items which each dentist could conceivably answer without direct reference to his patient-records, i.e., from personal knowledge, from impression or guess, from memory alone, or from records other than those dealing with treatment. The second part of the questionnaire consisted of six copies of a form on which the dentist was requested to list each patient-visit and the treatment and/or procedures performed during that visit. Each form was to contain the data for all patient-visits on a specific day during a stipulated week - the same week for all dentists and a week prior in time to actual receipt of the questionnaire. These data would, of course, have to be taken by the dentist from his patient-records.

The stipulated week turned out, unfortunately, to have been one during which some of the dentists in the study were attending a state dental convention. Partly for this reason, but clearly for other reasons as well, only about 30 percent of the questionnaires were returned, about 120 in all out of the about 400 listed in the Directory; the proportions returned were nearly identical for the fluoridated and fluoride-deficient communities. The entire group of dentists who returned their questionnaires answered all questions on the first part, but fewer than one-half of these (14 percent of all questionnaires sent out) gave any information at all in response to the second part of the questionnaire, i.e., on patient-visits and treatment and/or procedures performed. Very few of these gave complete data. These response rates were far below those obtained subsequently by personal interviews. The magnitude of these differences is so great that there can be little question of their reliability.

A related question to which the present study also addressed itself was this: Within the personal interview situation, was it better to ask the dentist to provide (or have an assistant provide) merely previously designated records, or even only to grant an interviewer access to the files from which the records could be pulled? The interviewer would then either photostat the designated records or he would abstract the required patient-care data from them. Or, should the dentist be asked to provide the required data already abstracted from the records? It was felt by the study staff that providing records, or granting access to the files containing these records, might be unacceptable to some dentists because of the confidentiality issue, while providing data might be unacceptable to others because of the time and effort required. However, there was no way of predicting what the proportions were likely to be in each of these categories, particularly since the public health and social science literature provided no guides whatever that might be used in answering these questions. The study staff was unable to find a single article reporting the use, as research data, of the records of dentists in private practice. Later, after the planning of the present study had been completed and its data gathered, Muhler [7] published some patient-care data obtained from dentists' records. (The usefulness of this report in planning the present study would have been sharply limited in any case, since it did not discuss any of the problems involved in obtaining these data, nor even how it was done.)

A complicating factor in planning this aspect of the study was that most dentists maintain more than one type of record containing at least some items of patient-care data. The most important among these types for present purposes is the patient-record itself, usually containing the profile of the patient's teeth; name, address, and (often) demographic data; history and clinical data often including a record of procedures, treatment, medications, and materials used during each visit; and financial data. Most dentists maintain, in addition, an appointment book, which usually specifies the reason for the visit, and many also use a day-sheet, or daily log-book. But in addition, some dentists now work with professional management firms, and these dentists may use daily billing sheets either in addition to or in place of day-sheets. The record of preference for the present type of research was clearly the patient-record itself.

In two feasibility surveys involving 102 dentists, about five out of six cooperating dentists (86 dentists, or 84 percent of the total) to whom these alternatives were presented - i.e., to provide an interviewer with records (or files containing these records), or to provide him with already abstracted data - were strongly reluctant to invest the time and effort necessary to abstract their records, or to have an assistant do so (see Table I). They much preferred to, and did, limit the extent of their responsibility in cooperating with the study only to granting access to their records or, at best, to pulling previously designated records from their files. In fact, many made it very clear that they would be unwilling to cooperate in the study if they (or their assistant) were expected to do the work necessary to provide

TABLE I. SOURCE OF PATIENT - CARE DATA (BY WHOM OBTAINED OR PROVIDED) AND TYPE OF RECORD USED

OFFICES OF 102 COOPERATING */ DENTISTS

Feasibility Surveys, Chicago and Champaign-Urbana

Type of Record Used	Source of Patient-Care Data		
	All sources	Abstracted by the interviewer	Abstracted by the dentist (or his assistant)
All records	102	86	16
Patient-record	78	71	7
Appointment book	6	3	3
Day-sheet	12	9	3
Billing sheet	4	3	1
Unknown	2	0	2

*/ Only dentists providing patient-care data. Dentists providing biographical and other data, but not patient-care data, are not included.

data rather than records. */ About one-sixth (16 dentists, or 14 percent of the total) provided data rather than permit the interviewer to have access to records or files.

However, only 71 of the 86 dentists willing to provide access to their records or files were willing to provide access to their patient-records, although this had been explicitly designated to them as the record of choice. Nine of the fifteen dentists were willing to permit access only to day sheets and three each to an appointment book and to billing sheets. Among the 16 dentists providing data rather than records, the work of abstracting the records was done in the interviewer's presence in 14 offices; **/ in one-half of these the data were obtained directly from patient-records.

Granted that the dentists much preferred not to be burdened with the work of abstracting the necessary data from their records, in favor of having it done by an interviewer, would this be feasible? Muller [8] has provided an apt analogy illustrating the difficulties involved in using records, although she qualifies her statement in the second sentence of the following quote:

> "In a sense the use of records for analytic study bears a resemblance to the work of the archaeologist reading the explicit and implicit messages in the residue of the past. But the social science investigator who studies prescribing has additional resources, such as direct observation, personal interview, and written questionnaires to amplify what is learned from prescribing records." ***/

*/ The mail survey, conducted as the initial phase of the present study, had in effect required participating dentists to provide data rather than records, thus committing them to the expenditure of much time and effort. This may have been a major factor in explaining the relatively poor response rate in that survey as compared, for example, to the relatively high rate obtained by Hochstim (better than 80 percent of questionnaires returned by mail) in a survey of households in Alameda County, California [4]. (However, another factor explaining Hochstim's much higher response rate was that he surveyed households, rather than professional offices as in the present survey.)

**/ The remaining two are accounted for as follows: In one case the abstracting was done prior to the interviewer's arrival, while in the other the data were subsequently mailed to the interviewer. In both of these instances the type of record serving as the source of the data remained unknown to the research staff.

***/ [8], p. 2117.

Muller's reference is to drug prescribing, and the special interest of investigators in this area, in the past at least, has been in drugs prescribed in a medical-care system in which there has been some type of third-party payer. Usually in these instances some standardization of forms or documents is involved, simply because of third-party payment; dentists in private practice, however, need use no such standardized forms, since usually no third-party payers are involved. In fact, they rarely expect anyone else to read their private patient-records; these records are usually written exclusively for their own use. */

The questions which follow immediately from this statement, therefore, and relating to the records themselves, are these: I) Do dentists use filing systems such that designated records can readily be found? 2) Once found, is the handwriting legible? 3) Do they use standard or personal idiosyncratic systems for numbering teeth? and 4) Do they use standard designations and abbreviations or symbols for dental procedures, treatments, medications and materials? The answers to these questions would determine whether it was, in fact, feasible to use patient-records. Two closely related questions, essentially aimed at how best to use these records, are these: I) Should the records be photostatted and brought to another setting where they could be interpreted with the aid of dental professional assistance as necessary? Or should they be abstracted by an interviewer in the dentist's office? If the latter course were found to be preferable, the next question is: 2) Would it be necessary, to interpret patient-records containing clinical data, that the interviewer have a great deal of technical training, e.g., that he be a dentist, a dental student, or a trained dental assistant or hygienist? Or could lay interviewers, even if previously unfamiliar with medical and/or dental terminology, be trained to do the job adequately? In a study of the office records of internists (physicians) in private practice, Altman [I] and his study staff first found it necessary to send a physician into another physician's office to judge the suitability of the latter's records for abstracting by a paramedically trained person; only then did the latter come in. However, medical records are apparently far

*/ This may not be altogether unlike the situation resulting in the difficulties which social researchers have experienced in attempting to motivate respondents to keep a diary of their expenditures over a period of time. See: Neter and Waksberg [9], pp. 1-2.

more complicated than dental records, and it did not appear to be necessary to follow the same procedure here.

Based on the feasibility surveys, it was very evident that dentists do have filing systems permitting ready access to designated records, and in general their handwriting was found to be legible. However, almost one-half of the dentists used personal, idiosyncratic systems for numbering teeth, rather than the Armed Forces or Standard Quadrant method taught in dental schools. Personal systems for designating procedures, treatments, medications, and materials were also quite common, and there was apparently no single system, universally in use among dentists, for abbreviation of clinical terms. */ The assistance of the dentist or a member of his staff in interpreting some portion of his patient-records was often essential and, as a result, it was found to be much preferable to abstract them in the dentist's office rather than having them photostatted and abstracted elsewhere. In addition, dentists were reluctant to have their records photostatted even with the names blocked out.

Despite the diversity among dentists in tooth-numbering systems and in designations and abbreviations of clinical terms, the feasibility surveys had suggested, and the main study confirmed this suggestion, that lay interviewers, even if previously unfamiliar with medical and/or dental terminology, could be trained within a relatively short period of time to do the job adequately if the training was sufficiently intensive. **/ The interviewers used in the study were ordinary interviewers of the type regularly employed by the National Opinion Research Center, without special prior background in medical or dental matters.

Finally, dentists' records in both the feasibility surveys and the main study proved to be relatively complete on clinical items and on purpose of visit (nearly all records had some notation on these matters), but relatively poor on age of patient. This was true of patient-records as well as of appointment books, day-sheets, etc. (The latter usually contain no age-data at all.) However, with regard to age, it was possible in a substantial number of instances to determine whether the patient was a child or an adult by reference to the patient-record containing his previous dental history, e.g., whether a deciduous tooth had been treated within a specified time-period prior to the survey week. While patients themselves would clearly have been a better source for this type of information, they nevertheless turned out, in one of the feasibility surveys, to be a relatively unreliable source of data on their own dental histories partly because of the memory factor. Patient-records were much better for this purpose, although a substantial limitation to these records was that they provided data only for the patient's current dentist; however, the memory factor does not pose a problem here.

* * * *

The above discussion reports some considerations relating to method in a pioneering study of the impact of fluoridation on dental practice. The bases for some important decisions are presented, and some limitations in study design and method are emphasized. Study of the patient-records of medical and dental practitioners in private practice is now widely recognized as an important key to understanding the structure and functions of the health services' system in this country, and it is hoped that, despite the difficulties inherent in this process, future studies will build upon the present endeavor.

*/ Judging by the above, communication errors are not likely to be confined to enumerative surveys involving verbal response. See: Mauldin and Marks [6].

**/ Full details of the training are available from Miss Coppersmith.

REFERENCES

[1] Altman, Isidore, "Assessing the quality of records in the physician's office." Presented at The Annual Meeting of the American Public Health Association, November 3, 1966. Processed.

[2] American Dental Association, Bureau of Economic Research and Statistics, The 1965 Survey of Dental Practice. Chicago: The Association, 1966.

[3] Douglas, Bruce L., and Coppersmith, Sylvia B., "The impact of water fluoridation on dental practice," New York State Dental Journal, 31 (1965), pp. 439-448.

[4] Hochstim, Joseph R., "A critical comparison of three strategies of collecting data from households," Journal of the American Statistical Association, 62 (1967), pp.976-989.

[5] Kirby, Joseph V., "Social and economic changes affecting dental practice," Journal of the American Dental Association, 58 (1959), pp. 84-91.

[6] Mauldin, W. Parker, and Marks, Eli S., "Problems of response in enumerative surveys," American Sociological Review, 15 (1950), pp. 649-657.

[7] Muhler, Joseph C., "The impact of preventive dentistry on dental practice," Journal of the American Dental Association, 74 (1967), pp. 111-113.

[8] Muller, Charlotte F., "The study of prescribing as a technic of examining a medical care system," American Journal of Public Health, 57 (1967), pp. 2117-2126.

[9] Neter, John, and Waksberg, Joseph, Response Errors in Collection of Expenditures Data by Household Interviews: An Experimental Study. Bureau of the Census, Technical Paper No. 11, Washington, D.C.: U.S. Government Printing Office, 1965.

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*

G. K. Kripalani**, Western Michigan University

1. The problem is to estimate for a color-sex-age group: (a) the number of net migrants for state i and (b) the net migration rate for state i subject to the following structural elements of the model:

(i) There are census enumeration errors.

(ii) U. S. population is subject to and is acted upon by only one cause of change viz. mortality.

(iii) State i population is subject to and is simultaneously acted upon by two causes of change viz. mortality and net migration (in- or out-).

2. Notation:

A = Number of persons in a color-sex-age (CSA) category, all aged "x" in U.S. as enumerated at the census at the beginning of a decade $t = 0$.

B = Number of persons in the same color-sex (CS) category, all aged "x + 10" in U.S. as enumerated at the census at the end of the decade $t = 1$.

A_i = Numbers of persons in the same CSA category, all aged "x" in state i as enumerated at the census at the beginning of the decade $t = 0$.

B_i = Number in the same CS category, all aged "x + 10" in state i as enumerated at the census at the end of the decade $t = 1$.

$R = B/A$ the census survival ratio for U.S. for the cohort.

$R_i = B_i/A_i$ the census survival ratio for state i for the same cohort.

A', B', A_i', B_i', denote corresponding 'true' unknown numbers.

3. A number of formulae using census survival ratios are currently in use for estimating internal net migration numbers and rates for individual color-sex-age groups in a state. Three of these formulae known as Forward, Reverse and Average formulae are given by Hamilton and Siegel (1942), and Hamilton (1959). The fourth formula is the one used by Everett S. Lee et al. (1957), in their momentous work, "Population Redistribution and Economic Growth, United States, 1870-1950"; and, the fifth is the formula which takes the numerator of the Forward formula for its numerator and the initial decade population for its denominator. These existing formulae are based on an arithmetical approach and thus provide single point estimates. The alternative method outlined below uses a statistical approach and hence provides an estimate of the standard error of the estimate of net migration rates and numbers. In addition further merits of the proposed method are:

(a) It calculates the net migration rate as an independent rate which excludes the effects of all other operating causes viz. mortality and census enumeration errors.

(b) Uses a procedure that takes into account the errors of differential under-enumeration and keeps such errors out of net migration rates and numbers. (While the estimates of net migration rate given by the three formulae of Siegel and Hamilton are free from the effects of census enumeration errors, <u>their estimates of the number of net migrants are not</u>.)

(c) Takes into account mortality differences between states and the national average.

(d) Provides an estimate of the measure of the impact of census enumeration errors considered as a separate cause acting on the model, besides the causes of net migration and mortality. (See paragraphs 10, 11 and 12)

---

* The major part of this work was done as a part of the author's work on a National Science Foundation Project "Area-Population Adjustments in Relation to Economic Activity" at North Carolina State University during 1963-65. The work was revised subsequently in 1967 when the author worked at the University of Chicago on a Corps of Engineers Project "Employment-Population Effects of Location of Water Resource Development Projects".

** Fellow, Institute of Actuaries, London.

4. The basic structural condition of our model is that a group of lives is being continuously and simultaneously acted upon by two causes of change viz. mortality and net migration, (ignoring, for the time-being the problem of census enumeration errors). The principal objection to all these formulae is that they are derived by a procedure that is valid only if the group of lives under consideration is subject to one cause of change at a time. This procedure is invalid when the group is subject to continuous and simultaneous operation of two or more causes of change. The basic structure of our problem is that we start with an initial group of lives at time $t = 0$ and that this group is being depleted by mortality and by net out-migration (or depleted by mortality and augmented by net in-migration) and that both causes of change operate simultaneously and continuously throughout the census decade up to time $t = 1$ (dropping for the time-being the problem of census enumeration errors). All the three formulae of Siegel and Hamilton assume that only one cause of change operated at a time. For example, the numerator of the Forward Formula represents the number of net migrants only if it is assumed that (a) mortality alone operates as a cause of change throughout the census decade and (b) at the end of the census decade for an instant of time, mortality ceases to operate and only net migration as a cause of change operates. This is the extreme kind of assumption underlying the Forward Formula. By a complete swing of the pendulum, as it were, the Reverse Formula is derived on the assumptions that net migration as a cause of change operates only for a brief instant at the start of the decade; thereafter throughout the decade, mortality alone acts as a cause of change affecting the group. Similarly the Average Formula is valid on the assumptions that (a) mortality alone operates as the cause of change from the start of the decade to the mid-point of the decade from $t = 0$ to $t = 1/2$, (b) precisely at $t = 1/2$ mortality ceases to operate and net migration as a cause of change operates for that instant and (c) mortality again takes over and operates as the sole cause of change from $t = 1/2$ to $t = 1$.

5. Some further comments are necessary on the Average Formula. In this formula, the numerator is simply the average of the numerators of the Forward and Reverse Formulae and so is the denominator. In fact, we could have an infinite number of Average Formulae by taking any weighted average of the numerators of the Forward and Reverse Formulae and corresponding weighted average of the denominators of the two formulae. If W_f is the weight attached to the numerator and the denominator of the Forward Formula and $W_r = 1 - W_f$ is the weight attached to the Reverse Formula, any weighted average formula obtained by the use of these weights will still give the same net migration rate. If the Forward Formula for the net migration rate is N_f/D_f and the Reverse Formula is N_r/D_r, then the generalized weighted Average Formula will be

$$\frac{W_f \cdot N_f + W_r \cdot N_r}{W_f \cdot D_f + W_r \cdot D_r} \quad . \; . \; . \; . \; . \; . \; . \; . \; . \; . \quad (1.7.1)$$

and will be equal to N_f/D_f or N_r/D_r since we have N_f/D_f equal to N_r/D_r. The generalized formula will hold for all arbitrary values of W_f or W_r lying in the closed interval 0 to 1. Thus, we can use any value for the numerator of the expression for the net migration rate lying between N_f and N_r provided we adjust the denominator accordingly. In fact, one may look at the Forward and Reverse Formulae as the special cases of the generalized formula when $W_f = 1$ and $W_r = 1$ respectively.

6. What passes as three formulae for the net migration rate is really a single formula valid for a distorted structure of the model in which mortality operates as the only cause of decrement throughout the closed interval of time $0 \leq t \leq 1$ except at a point of time $t = t^{**}$ and in which net migration as a cause affecting the group does not operate at all throughout the decade except at the brief instant of time $t = t^{**}$. As the generalized weighted average formula shows the choice of t^{**} is absolutely arbitrary and does not affect the value of the ratio viz. the net migration rate.

7. To summarize, the main objection to the procedure underlying the derivation of the three formulae of Siegel and Hamilton is that it is valid only when the group of lives is subject to one cause of change at a time and not when the structural situation is one in which the group of lives is continuously and simultaneously acted upon by two causes of change. Secondly, in terms of the structure of the model, neither the numerator nor the denominator of any of the formulae, possesses real interpretation or significance. Thirdly, none of the expressions for the number of net migrants given by Siegel and Hamilton is free from the effect of census enumeration errors. Similar comments apply to formulae (4 and 5).

8. Notation:

Let p^d and p_i^d denote the independent survival rate against mortality for the specified CSA category during the decade, for U. S. and state i respectively. Hence $q^d = 1 - p^d$ and $q_i^d = 1 - p_i^d$ where q^d and q_i^d are independent mortality rates.

Let p_i^w denote the independent survival rate against net migration for state i for the specified CSA category during the decade. Hence $q_i^w = 1 - p_i^w$ where q_i^w is the independent rate of net migration. Taking net in-migration as negative and net out-migration as positive,[1] we have when there is net in-migration $p_i^w > 1$ since q_i^w is negative, and when there is net out-migration $p_i^w < 1$ since q_i^w is positive.

9. It will be assumed that: (a) the U.S. population (numbering A at the start of the decade) is a random sample from an infinitely large population for which probability of dying in the unit time interval (defined as an inter-censal decade) is $Q = 1 - P$ where P is the Life Table (L.T.) 10-year Survival Rate for the particular CSA category

in U. S. population for the relevant intercensal period; and (b) state i population (number A_i at the start of the decade) is a random sample from an infinitely large population for which probability of dying in the unit time interval (defined as an intercensal decade) is $Q_i = 1 - P_i$ where P_i is the Life Table (L.T.) 10-year Survival Rate for the particular CSA category in state i population for the relevant intercensal period.

10. It is important at this stage to comment upon the manner in which the significance of the impact of census enumeration errors is proposed to be viewed. For U. S., $A = (1 - e_o)A'$ and $B = (1 - e_1)B'$ so that $R = R' \times (1 - e_o) / (1 - e_1) = KR'$ where $K = (1 - e_o) / (1 - e_1)$, and e_o and e_1 represent the extent of under-enumeration in U. S. at census at t = 0 and t = 1 respectively. Similarly for state i, $R_i = K_i R_i'$ where $K_i = (1 - e_o^i) / (1 - e_1^i)$ and e_o^i and e_1^i represent extent of underenumeration in state i at t = 0 and t = 1 respectively.

It may be observed that K (or K_i) is always positive. Whenever there is underenumeration e's are positive fractions $0 \leq e < 1$, so that the factors (1 - e) are always positive. When there is over-enumeration, corresponding e is negative so that $1 - e > 1$ is always positive.

11. Let us define η_t by the relationship

$$K = \exp\left[-\int_0^1 \eta_t \, dt\right] = e^{-\eta} \quad \ldots\ldots\ldots\ldots (1)$$

where $\eta = \int_0^1 \eta_t \, dt$.

$$\text{or } \eta = -\log_e K \quad \ldots\ldots\ldots\ldots\ldots\ldots (2)$$

$$\eta \gtreqless 0 \text{ according as } K \lesseqgtr 1 \quad \ldots\ldots\ldots (3)$$

When $K > 1$, $\eta < 0$ so that $-\eta > 0$. In that case, let $\delta = -\eta > 0$, so that $K (>1) = e^{\delta}$ where $\delta > 0$.

When K = 1, the differential errors of underenumeration have no effect on the census survival ratio and the observed CSR is the same as true CSR. When K < 1, the effect of the errors of differential underenumeration is equivalent to the operation of a rate of decrement whose average force of exit over the intercensal decade (interval 0, 1) is $\eta = -\log_e K$. When K > 1, the effect is equivalent to the operation of a rate of increment whose average force of entry, is $\delta = \log_e K$. In general

$$R = KR' = R' e^{-\int_0^1 \eta_t \, dt} = R' e^{-\eta} \ldots (4)$$

Similar remarks apply to K_i for state i. We have

$$\frac{B_i}{A_i} = K_i \frac{B_i'}{A_i'}$$

and $K_i = e^{-\int_0^1 \eta_{it} \, dt} = e^{-\eta_i}$

12. We may thus view the situation for each CSA category in U.S. population as enumerated as a case of a population subject to the operation of two causes, (i) mortality and (ii) a cause of exit (or entry) whose average force of exit (or entry) is η so that the independent 'survival' rate against this cause is $e^{-\eta} = K$. Similarly we will view the situation for each CSA category in a state i population as enumerated as a case of a population subject to the operation of three causes, (i) mortality (ii) net migration (in or out) and (iii) a cause of exit (or entry) whose average force of exit (or entry) is η_i so that the independent 'survival' rate against this cause is $e^{-\eta_i} = K_i$.

13. The proposed method is based on a fundamental theorem in the Theory of Life Contingencies which states that "when a group of lives is simultaneously acted upon by two or more independent causes of change[2], the over-all rate of survival against all causes acting together is equal to the product of the various rates of survival against individual causes acting separately. Thus if a group of lives aged 'x' is subject to the simultaneous operation of three independent causes α, β and γ, then (in actuarial notation):

$$ap_x = p_x^{\alpha} \cdot p_x^{\beta} \cdot p_x^{\gamma}$$

where ap_x denotes the over-all rate of survival against all causes combined and p_x^{α} denotes the independent rate of survival against cause α alone.

14. Assuming that the effect of census enumeration errors on the observed census survival ratio is the same for U. S. as for each state i for a specified CSA category, i.e., $K = K_i$, (This assumption is the same as has been made by Zachariah (1962) and others.) We have:

---

[1]Usually the present formulas treat net out-migration as negative and net in-migration as positive. In this paper, net out-migration has been handled as positive and net in-migration as negative. The reason for doing so is that if net out-migration rate (say, denoted by q) is treated as positive, p = 1 - q acquires a real physical significance, and represents the survival rate against the cause of "net migration" in the same way as when q is the life table mortality rate, p = 1 - q represents the life table survival rate. If net out-migration is treated as negative as is done at present by Hamilton & Siegel and others, p = 1 - q > 1 will have no significance whatsoever. Thus, the conventions set up in this paper of treating net out-migration as positive and net in-migration as negative, give real significance to the complementary quantity p = 1 - q and enable the use of binomial distribution when q > 0 (net out-migration) and negative binomial when q < 0 (net in-migration), p < 1 in the case of net out-migration and p > 1 in the case of net in-migration.

For U. S. $R = K \cdot p^d$ ....(5)
For state i $R_i = K\, p_i^d\, p_i^w$ ....(6)

on the basis of the fundamental formula

$$ap_x = p_x^{\alpha}\, p_x^{\beta}\, p_x^{\gamma}.$$

15. In terms of the assumptions of para 9., we may regard p^d as a binomial variable with mean P and variance PQ/A[3] where A is the initial population in the specified CSA category in U. S. at the start of the decade. Similarly p_i^d may be regarded as a binomial variable with mean P_i and variance P_iQ_i/A_i

$$p^d \sim \text{Bin}\ (P,\ PQ/A) \ \ldots\ldots\ldots\ldots (7)$$

$$p^d \sim \text{Bin}\ (P_i,\ P_iQ_i/A_i) \ \ldots\ldots\ldots\ldots (8)$$

Since A and A_i are generally very large, p^d and p_i^d may be regarded as distributed normally.

16. Let us write:

For U.S.: $R = Kp^d = KP\varepsilon_o$ ....(9)
where ε_o is a random variable.

For state i: $R_i = Kp_i^d\, p_i^w = Kp_i^w\, P_i\, \varepsilon_i$ (10)
where ε_i is a random variable.

From (9) we have

$$p^d = P\ \varepsilon_o$$
$$E\ (p^d) = PE\ (\varepsilon_o)$$

But $E\ (p^d) = P$, so that

$$E\ (\varepsilon_o) = 1 \ \ldots\ldots\ldots\ldots (11)$$
$$V\ (p^d) = V\ (P\ \varepsilon_o)$$
$$\frac{P(1-P)}{A} = P^2\ V\ (\varepsilon_o)$$
$$\text{or } V\ (\varepsilon_o) = \frac{1-P}{A\ P} \ \ldots\ldots\ldots\ldots (12)$$

[2]The theorem is generally proved in any text book on "Life Contingencies" in relation to causes of decrement, but it can easily be established in the more general case when the causes operating are the causes of change, some or all of which may be of incremental type. The proof is briefly as follows:

$$ap_x = \exp\left[-\int_0^1 \mu_{x+t}\, dt\right]$$

where μ_{x+t} is the over-all force of decrement (or change). $p_x^{\alpha} = \exp\left[-\int_0^1 \mu_{x+t}\, dt\right]$ etc.

If the forces of decrement (or change) due to causes α, β and γ are independent $\mu_{x+t}^{\alpha} + \mu_{x+t}^{\beta} + \mu_{x+t}^{\gamma}$. Hence the result.

[3]Strictly speaking, the denominators in the expressions for variances are not A or A_i but the true exposed to the risk of death.

Similarly,

$$E\ (\varepsilon_i) = 1 \ \ldots\ldots\ldots\ldots (13)$$
$$V\ (\varepsilon_i) = \frac{1-P_i}{A_iP_i} \ \ldots\ldots\ldots\ldots (14)$$

17. Note that p^d and p_i^d are independent and hence ε_o and ε_i are independent. Further the variability of $\varepsilon_o = \frac{\sigma(\varepsilon_o)}{E(\varepsilon_o)} = \sqrt{\frac{1-P}{AP}}$ is generally near to 1 and A is very large. We may therefore, use the following results for the mean value and variance of a ratio:

$$E\left(\frac{X_1}{X_2}\right) = \frac{E(X_1)}{E(X_2)} \ \ldots\ldots\ldots\ldots (15)$$

$$V\left(\frac{X_1}{X_2}\right) = \frac{\xi_1^2}{\xi_2^2}\left[\frac{\sigma_1^2}{\varepsilon_1^2} + \frac{\sigma_2^2}{\varepsilon_2^2} + \frac{3\sigma_1^2\ \sigma_2^2}{\varepsilon_1^2\ \varepsilon_2^2}\right] \ldots\ldots (16)$$

where $E(X_1) = \xi_1$; $E(X_2) = \xi_2$; $V(X_1) = \sigma_1^2$, $V(X_2) = \sigma_2^2$

and (i) X_1 and X_2 are independent; and

(ii) $\frac{\sigma_2}{\xi_2}$ is small so that its higher than second powers can be ignored. We can therefore take:

$$E\left(\frac{\varepsilon_i}{\varepsilon_o}\right) = \frac{E(\varepsilon_i)}{E(\varepsilon_o)} = 1 \ \ldots\ldots\ldots\ldots (17)$$

Now
$$p_i^w = \frac{R_i}{R}\ \frac{P}{P_i}\ \frac{\varepsilon_i}{\varepsilon_o}$$

$$E(p_i^w) = \frac{R_i}{R}\ \frac{P}{P_i}\ E\left(\frac{\varepsilon_i}{\varepsilon_o}\right) = \frac{R_i}{R}\ \frac{P}{P_i} \ \ldots\ldots\ldots (18)$$

We may take $\frac{R_i}{R}\ \frac{P}{P_i}$ as an unbiased estimate of p_i^w.

$$\hat{p}_i^w = \frac{R_i}{R} \cdot \frac{P}{P_i}$$

$$\hat{q}_i^w = 1 - \hat{p}_i^w = 1 - \frac{R_i}{R} \cdot \frac{P}{P_i} \ \ldots\ldots\ldots\ldots (19)$$

Estimate of the independent rate of migration is given by $1 - \frac{R_i}{R} \cdot \frac{P}{P_i}$.

18. We have already shown that,

$E(\varepsilon_o) = 1$, $V(\varepsilon_o) = \frac{1-P}{AP}$. Denote by σ^2.

$E(\varepsilon_i) = 1$, $V(\varepsilon_i) = \frac{1-P_i}{A_iP_i}$. Denote by σ_i^2.

As in our problem $\frac{\sigma(\varepsilon_o)}{E(\varepsilon_o)}$ and $\frac{\sigma(\varepsilon_i)}{E(\varepsilon_i)}$ are both small, we apply the approximate relationship (16) and omit the last term in the bracket viz.

$$\frac{3\sigma_1^2\sigma_2^2}{\xi_1^2\xi_2^2} .$$

We have

$$V\left(\frac{\varepsilon_i}{\varepsilon_o}\right) = V(\varepsilon_o) + V(\varepsilon_i) + 3V(\varepsilon_o) \, . \, V(\varepsilon_i). \quad (20)$$

$$= \sigma^2 + \sigma_i^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (21)$$

ignoring the last term.

$$V(p_i^w) = V\left(\frac{R_i}{R}\,\frac{P}{P_i}\,\frac{\varepsilon_i}{\varepsilon_o}\right)$$

$$= \frac{R_i^2}{R^2}\,\frac{P^2}{P_i^2}\left[\frac{Q}{AP} + \frac{Q_i}{A_iP_i}\right] \quad \ldots\ldots\ldots\ldots\ldots\ldots (22)$$

Where $Q - 1 - P$ and $Q_i = 1 - P_i$.

19. $q_i^w = 1 - p_i^w$ and $\hat{q}_i^w = 1 - \hat{p}_i^w$

$\hat{q}_i^w > 0$ when $\hat{p}_i^w > 1$ or there is out-migration.

$\hat{q}_i^w < 0$ when $\hat{p}_i^w > 1$ or there is in-migration.

$$E\,(q_i^w) = E(1 - p_i^w) = 1 - \frac{R_i}{R}\,\frac{P}{P_i} \quad \ldots\ldots\ldots\ldots (23)$$

$$V(q_i^w) = V(1 - p_i^w) = V(p_i^w)$$

$$= \frac{R_i^2P^2}{R^2P_i^2}\left[\frac{Q}{AP} + \frac{Q_i}{A_iP_i}\right] \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (24)$$

20. To estimate the expected number of net migrants ad_i^w, we will use the following formula, from the Theory of Life Contingencies:

$$ad_i^w = A_iq_i^w \left[1 - 1/2\,(q_i^d + q_i^k) + 1/3\, q_i^d \, . \, q_i^k\right] \quad (25)$$

where

$$q_i^d = 1 - p_i^d \text{ and } q_i^k = 1 - K$$

Substituting for p_i^d and K, we have:

$$ad_i^w = 1/6\,A_i \left[2 + P_i\varepsilon_i + \frac{R}{P}\,\frac{1}{\varepsilon_o} + 2\,\frac{P_iR}{P}\,\frac{\varepsilon_i}{\varepsilon_o} - 2\,\frac{R_i}{R}\,\frac{P}{P_i}\,\frac{\varepsilon_o}{\varepsilon_i} - \frac{R_iP}{R}\,.\,\varepsilon_o - \frac{R_i}{P_i}\,\frac{1}{\varepsilon_i} - 2\,R_i\right] \ldots\ldots (26)$$

Using the results:

$$E(\varepsilon_i) = E(\varepsilon_o) = 1 \qquad E\left(\frac{\varepsilon_i}{\varepsilon_o}\right) = \frac{E(\varepsilon_i)}{E(\varepsilon_o)} = 1$$

$E\left(\frac{1}{\varepsilon_o}\right) = 1 \quad E\left(\frac{1}{\varepsilon_i}\right) = 1$, it can be shown that

$$E(ad_i^w) = \frac{A_i}{6}\left[1 - \frac{R_i}{R}\,\frac{P}{P_i}\right]\left[2 + P_i + \frac{R}{P} + 2\,R\,.\,\frac{P_i}{P}\right] \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (27)$$

21. Let us rearrange the terms in ad_i^w from (26)

$$ad_i^w = \frac{A_i}{6}\left[2 - 2\,R_i + P_i\varepsilon_i - \frac{R_iP}{R}\varepsilon_o - \frac{R_i}{P_i}\,\frac{1}{\varepsilon_i} + \frac{R}{P}\,\frac{1}{\varepsilon_o} + 2\,\frac{P_iR}{P}\,\frac{\varepsilon_i}{\varepsilon_o} - 2\,\frac{R_iP}{RP_i}\,\frac{\varepsilon_o}{\varepsilon_i}\right]$$

$$= \frac{A_i}{6}\left[2 - 2\,R_i + C_1\varepsilon_i + C_2\varepsilon_o + C_3\,\frac{1}{\varepsilon_i} + C_4\,\frac{1}{\varepsilon_o} + C_5\,\frac{\varepsilon_i}{\varepsilon_o} + C_6\,\frac{\varepsilon_o}{\varepsilon_i}\right] \quad \ldots\ldots\ldots\ldots\ldots\ldots (28)$$

$$V(ad_i^w) = \frac{A_i^2}{36}\left[V\left(C_1\varepsilon_i + C_2\varepsilon_o + C_3\,\frac{1}{\varepsilon_i} + C_4\,\frac{1}{\varepsilon_o} + C_5\,\frac{\varepsilon_i}{\varepsilon_o} + C_6\,\frac{\varepsilon_o}{\varepsilon_i}\right)\right] = \frac{A_i^2}{36}\left[C_1^2\,V(\varepsilon_i) + C_2^2\,V(\varepsilon_o) + C_3^2\,V\left(\frac{1}{\varepsilon_i}\right) + C_4^2\,V\left(\frac{1}{\varepsilon_o}\right) + C_5^2\,V\left(\frac{\varepsilon_i}{\varepsilon_o}\right) + C_6^2\,V\left(\frac{\varepsilon_o}{\varepsilon_i}\right) + \text{Covariance Terms}\right] \ldots\ldots\ldots\ldots (29)$$

There will be $C_2^6 = 15$ convariance terms whose coefficients will be $2C_iC_j$ $i \neq j$. $(i = 1, 2, \ldots 6,\ j = 1, 2, \ldots ¢)$. Of these, 11 of the covariance terms are all zero when we use the independence of ε_o and ε_i and use the results $E\left(\frac{1}{\varepsilon_i}\right) = \frac{1}{E(\varepsilon_i)}$ etc. These terms correspond to following C_iC_j coefficients.

C_1C_2; C_1C_3; C_1C_4; C_1C_6; C_2C_3; C_2C_4; C_2C_5; C_3C_4; C_3C_5; C_4C_6; C_5C_6

The non zero convariance terms are:

$\text{Cov}\left(\varepsilon_i, \frac{\varepsilon_i}{\varepsilon_o}\right) = V(\varepsilon_i)$ Coefficient $2\,C_1C_5$

$\text{Cov}\left(\varepsilon_o, \frac{\varepsilon_o}{\varepsilon_i}\right) = V(\varepsilon_o)$ Coefficient $2\,C_2C_6$

$\text{Cov}\left(\frac{1}{\varepsilon_i}, \frac{\varepsilon_o}{\varepsilon_i}\right) = \frac{-Q_i}{Q_i + A_iP_i} = \frac{-V(\varepsilon_i)}{1 + V(\varepsilon_i)}$, coefficient $2\,C_3C_6$ ....(30)

$\text{Cov}\left(\frac{1}{\varepsilon_o}, \frac{\varepsilon_i}{\varepsilon_o}\right) = \frac{-Q}{Q + AP} = \frac{-V(\varepsilon_o)}{1 + V(\varepsilon_o)}$, coefficient $2\,C_4C_5$

Dealing with the variance terms, we have by use of (16)

$$V\left(\frac{1}{\varepsilon_i}\right) = V(\varepsilon_i) = \sigma_i^2$$

$$V\left(\frac{1}{\varepsilon_o}\right) = V(\varepsilon_o) = \sigma^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (31)$$

$$V\left(\frac{\varepsilon_o}{\varepsilon_i}\right) = V\left(\frac{\varepsilon_i}{\varepsilon_o}\right) = V(\varepsilon_o) + V(\varepsilon_i) = \sigma^2 + \sigma_i^2$$

Going back to (29) and noting that σ^2 and σ_i^2 are very small compared to 1 so that we can use the approximate formula $(1 + X)^{-1} = 1 - X$ by ignoring second and higher powers of X. On this assumption $X(1 + X)^{-1} = X$. Therefore if we ignore σ^4 and σ_i^4 as being negligible, it can be shown that

$$V(ad_i^W) = \frac{A_i^2}{36} \left[\sigma_i^2 \left\{ P_i^2 + \frac{R_i^2}{P_i^2} + 4P_i^2 \frac{R^2}{P^2} + \frac{4R_i^2 P2}{R^2 P_i^2} + 4P_i^2 \frac{R}{P} - 4 \frac{R_i^2}{P_i^2} \frac{P}{R} \right\} + \sigma^2 \left\{ \frac{R_i^2}{R^2} P^2 + \frac{R^2}{P^2} + 4P_i^2 \frac{R^2}{P^2} + \frac{4R_i^2 P^2}{R^2 P_i^2} + \frac{4R_i^2 P2}{R^2 P_i} - \frac{4R^2}{P^2} \quad P_i \right\} \right]$$

Let us write:

$$\bar{W}_i = \frac{R_i}{R} \cdot \frac{P}{P_i} \text{ and } \hat{K} = \frac{R}{P} \text{ and } \lambda_i = \frac{R_i}{P_i}$$

so that $\bar{W}_i = \lambda_i / \hat{K}$.

In our expression we retain P, P_i, $\hat{K}$ and $\bar{W}_i$.

$$V(ad_i^W) = \frac{A_i^2}{36} \left[\sigma_i^2 \left\{ P_i^2 (1 + 2\hat{K})^2 + \bar{W}_i^2 (2 - \hat{K})^2 \right\} + \sigma^2 \left\{ \hat{K}^2 (2P_i - 1)^2 + \bar{W}_i^2 (2 + P_i)^2 \right\} \right] \quad ..(32)$$

22. By way of illustration, necessary calculations have been done for the state of North Carolina for 1950-60 decade in respect of white males (Appendices A and B). By way of comparison, the estimates of the number of net migrants and of net migration rates given by the proposed method and by U.S. Department of Agriculture for 1950-60 decade for the state of North Carolina for white male and white female categories are shown in Appendices C and D.

LIST OF REFERENCES

Bailey, W. G., and Haycocks, H. W., "Some Theoretical Aspects of Multiple Decrement Tables," (1946) Institute of Actuaries, London.

Hamilton, C. Horace, "Educational Selectivity of Net Migration from the South," Social Forces 38 (1959) 33-42.

Hamilton, C. Horace, and Henderson, F. M., "Use of the Survival Rate Method in Measuring Net Migration," Journal of the American Statistical Association, 39 (1944) 197-206.

Lee, Everett S., and Lee, Anne S., "Internal Migration Statistics for the United States," Journal of the American Statistical Association, 55 (1960) 664-697.

Lee, Everett S., et al., "Population Redistribution and Economic Growth, United States, 1870-1950," Volume 1. Philadelphia: The American Philosophical Society, 1957.

Price, Daniel O., "Examination of Two Sources of Error in the Estimation of Net Internal Migration," Journal of the American Statistical Association, 50 (1955) 689-700.

Siegel, Jacob S., and Hamilton, C. Horace, "Some Considerations in the Use of the Residual Method of Estimating Net Migration," Journal of the American Statistical Association, 47 (1952) 475-500.

Spuregeon, "Life Contingencies," Institute of Actuaries, London.

Wolfenden, Hugh H., "Population Statistics and Their Compilation, (1954)," published for the Society of Actuaries by the University of Chicago Press.

Zachariah, K. C., "A Note on the Census Survival Method of Estimating Net Migration," Journal of the American Statistical Association, 57 (1962) 175-183.

APPENDIX A

Main results of Calculations pertaining to:

(a) Expected Value and
(b) Variance of net out-migration rate

(i) North Carolina--White Male--1950-60 Decade

Age in 1950 X	Formula 5@	PROPOSED METHOD	
	$(MR)_i^u$	$E(q_i^W)$	$V(q_i^W)$
(1)	(2)	(3)	(4)
			10^{-5}
0-4	.03420	.03312	.00409
5-9	-.02867	-.02964	.00550
10-14	.02670	.02814	.00889
15-19	.13318	.13536	.00953
20-24	.10822	.10580	.01182
25-29	.06193	.05758	.01696
30-34	.03960	.03511	.02907
35-39	.04019	.03361	.04810
40-44	.02158	.01311	.09142
45-49	.01373	.00883	.17300
50-54	.01997	.01916	.30940
55-59	-.01575	-.01644	.62533
60-64	-.02620	-.01503	1.19502
65+	-.00089	.01890	1.93990

@$(MR)_i^u = (A_i R - B_i) / A_i = R - R_i$.

(Note: A negative value signifies net in-migration rate)

APPENDIX B

Main results of calculations pertaining to:

(a) Expected Value and
(b) Variance of the number of net out-migrants

(i) North Carolina--White Male--1950-60 Decade

Age in 1950 (X)	Forward formula[@] Net out-migrants $(NM)_i^u$	PROPOSED METHOD $E(ad_i^w)$	PROPOSED METHOD $V(ad_i^w)$
(1)	(2)	(3)	(4)
0-4	6,059	5,910	361
5-9	- 4,349	- 4,459	324
10-14	3,584	3,692	441
15-19	17,756	17,900	576
20-24	14,350	14,042	676
25-29	7,964	7,372	841
30-34	4,524	3,980	1,089
35-39	4,325	3,526	1,521
40-44	1,990	1,159	1,936
45-49	1,043	639	2,304
50-54	1,298	1,137	2,704
55-59	- 839	- 794	3,136
60-64	- 1,108	- 555	3,136
65+	- 73	1,124	8,649

[@] $(NM)_i^u = A_iR - B_i$.

(Note: A negative value signifies net in-migrants)

APPENDIX C

North Carolina--1950-60 Decade--Net Migration Rate Estimates Given by the USDA and by the Proposed Method

Age 1960	White Male USDA	White Male Proposed Method = P.M.	White Female USDA	White Female P.M.
10-14	.032	.033	.037	.038
15-19	-.031	-.030	.023	.024
20-24	.026	.028	.064	.074
25-29	.134	.135	.091	.093
30-34	.106	.106	.070	.071
35-39	.060	.058	.045	.047
40-44	.038	.035	.031	.033
45-49	.040	.034	.021	.024
50-54	.020	.013	.004	.010
55-59	.011	.009	.002	.009
60-64	.019	.019	-.007	.004
65-69	-.026	-.016	-.023	-.005
70-74	-.045	-.015	-.020	.005
75+	-.010	.019	-.009	.025

Source of USDA Estimates: Net Migration of the Population, 1950-60, by Age, Sex and Color, Volume 1, Part 3, page 459.

APPENDIX D

North Carolina--1950-60 Decade--Number of Net Migrants Estimates Given by USDA and the Proposed Method

Age 1960	White Male USDA	White Male P.M.	White Female USDA	White Female P.M.
10-14	5802	5910	6454	6509
15-19	-4604	-4459	3337	3543
20-24	3327	3692	8189	9606
25-29	17499	17900	11699	11889
30-34	14097	14042	9122	9331
35-39	7712	7372	5863	6095
40-44	4272	3980	3534	3780
45-49	4071	3526	2199	2599
50-54	1734	1159	386	901
55-59	780	639	159	729
60-64	1040	1137	- 407	258
65-69	-1099	- 794	-1192	- 287
70-74	-1370	- 555	- 752	223
75+	- 336	1124	- 417	1841

Source of USDA Estimates: Net Migration of the Population, 1950-60, by Age, Sex and Color, Volume 1 Part 3, page 459.

WELFARE RULES AND THE UNDERENUMERATION OF NONWHITE MALES

Nancy Jacoby and Ralph Novoa
Bureau of the Census

Summary

The purpose of this study was to investigate the possible relationship between "strict" State welfare rules and the undercount of young adult nonwhite males. Although some of the results of the study are not inconsistent with the hypothesis that States with "strict" welfare rules will have a greater undercount of nonwhite males than will States with "nonstrict" welfare rules, the limitations of the study are serious enough to prohibit any definite conclusions.

Background[1]

Recently, much attention has been focused on underenumeration in the census and its possible causes. Demographic analysis has provided estimates of the magnitude of the errors in the census counts, showing that a disproportionate share of the undercount consisted of young adult nonwhite males. Deliberate concealment is probably one of the factors in underenumeration, but just how great a factor it is, is not known. A person may be concealed for any of a number of reasons, and one possible reason involves welfare rules. One theory that has been proposed is that welfare rules that deny funds to families because of an employable but unemployed person in the household or because of an illicit living arrangement may cause individuals (especially males) not to be reported in the census.

Under the Social Security Act, Federal funds are available to States for their programs of old-age assistance, aid to families with dependent children (AFDC), aid to the blind, and aid to the permanently and totally disabled. If a State wishes to receive Federal funds, it must have a plan which is approved by the Department of Health, Education, and Welfare as meeting the requirements set forth in the Social Security Act. The general requirements for AFDC state that funds are available for assistance of children "deprived of parental support or care by reason of death, continued absence from home, or physical or mental incapacity of a parent." However, several States have an additional requirement, the "substitute parent" requirement, which is an outgrowth of efforts to define "deprivation of parent support" and "needy child" so as to exclude certain families who would normally be considered eligible, without invoking federal sanctions. In these States a man who has relationship with a mother on welfare, whatever its quality or duration, is considered a "substitute parent" and the children are not considered to be "deprived of parental support" or "needy."

Another type of public assistance is general assistance. This program is financed completely from State and/or local funds and is available primarily to those people in need who cannot qualify for help under one of the four State-Federal programs. There are wide variations in the content and application of the general assistance program from State to State, but in 15 States in 1960 there was a requirement that there could be no employable person in the family for the family to receive funds.

Although all States have welfare rules which could encourage concealment, certain States have particularly stringent rules which might increase the likelihood of individuals being concealed. These "strict rule" States are the States which have either a "substitute parent" type requirement for AFDC, or an employable person rule for general assistance, or both. Because of the subtle differences between the rules in some States, classification of the States as to stringency of rules becomes somewhat arbitrary. For this reason several lists of "strict rule" States were formed based on various sources of information.

Preparation of Lists[1]

Because of the small percentage of Negroes in Hawaii and Alaska, these States were omitted from the five lists of "strict rule" and "nonstrict rule" States which were formed for this investigation. The five lists were formed as follows:

List 1: A list based on intensive research into various welfare laws in effect in April, 1960. Alabama, Arkansas, Florida, Georgia, Louisiana, Michigan, Mississippi, Oregon, South Carolina, Texas, and Virginia had a "substitute parent" type requirement in their welfare policies in 1960. Arizona, D. C., Georgia, Iowa, Louisiana, Mississippi, Missouri, Nevada, North Carolina, Oklahoma, South Carolina, Texas, and West Virginia required that there be no employable persons in the family for the family to receive funds. These States were designated as "strict rule" States; the remaining States with the exception of Maryland were designated as "nonstrict rule" States. Maryland was classified as "unknown" because her welfare rules were not consistent throughout the State.

List 2: A list of 8 "strict rule" and 31 "nonstrict rule" States compiled by an expert in the field of child welfare at the Department of Health, Education, and Welfare. The remaining States were classified as "unknown."

List 3: A list composed of the 8 "strict rule" and 31 "nonstrict rule" States in list 2 plus two additional "strict rule" States (Arkansas and Texas). These 2 States probably but not definitely had "strict" welfare rules in 1960. The remaining States were classified as "unknown."

[1] Much of the information on welfare policies was obtained from Bell, Winifred, "Aid to Dependent Children." Columbia University Press, N.Y., 1965.

[1] See Appendix A for detailed composition of each list.

List 4: A list of the 32 States which were in agreement on lists 1 and 2. The remaining States were classified as "unknown."

List 5: A list of the 34 States which were in agreement on list 1 and 3. The remaining States were classified as "unknown."

After the lists were formed, various statistics which would be affected by an undercount were gathered for the nonwhite population. These statistics were then compared for the "strict rule" and "nonstrict rule" States on each list.

Results

The primary statistic analyzed in this investigation was the sex ratio for nonwhites.[1] Table 1 shows median sex ratios for "strict rule" and "nonstrict rule" States, for total U. S.

A comparison of sex ratios for States with "strict" and "nonstrict" welfare rules shows that the "nonstrict rule" States have significantly higher sex ratios (at the .05 level) for nonwhites in the age group 20-49 than do the "strict rule" States (see Table 2).[2] However, this same relationship is found for these groups of States when data from the 1910 Census are analyzed (before welfare rules were prevalent).

The result of Table 2 indicates that States with "strict" welfare rules probably have other factors influencing the sex ratios besides deliberate concealment due to welfare rules.

Since the States with "strict welfare rules are concentrated in the South region, it is important to compare the "strict rule" and "nonstrict rule" States for just the South. Table 3 shows median sex ratios for "strict rule" and "nonstrict rule" States for the South.

A comparison of the States shows only one significant difference in sex ratios using data from the 1960 Census (see Table 4). Data from the 1910 Census produces a similar result. The small number of States involved in the within-South comparison reduces the ability to detect differences.

There are probably other factors affecting the State data which have a more serious distorting effect on sex ratios than does concealment. One of the most important of these factors is migration. Mobility studies indicate that a greater number of nonwhite males than nonwhite females emigrate from the South every year. Migration of nonwhite males of employable age from the South reduces sex ratios the same as would concealment. The comparison of the 1960 data with the 1910 data shows that if concealment is a partial explanation for missed persons in the 1960 Census, the extent cannot be determined from available data.

Additional comparisons were made (using 1960 Census data) of sex ratios of nonwhites between the ages of 20 and 39 in central cities. Table 5 shows median sex ratios for "strict rule" and "nonstrict rule" States, for central cities.

The results of the comparisons were similar to those obtained for nonwhites between the ages of 20 and 49 on the State level (see Table 6).

Another statistic tested was the proportion of nonwhite households with female heads in 1960 in the central cities of each State. The test results showed no significant differences between "strict rule" States and "nonstrict rule" States for this statistic.[1]

Additional tables, not presented in this report, showing selected characteristics of the nonwhite population in central cities of SMSA's by State and by South Atlantic and South Central regions, 1960 Census are also available.

Limitations

Several limitations to this study must be kept in mind: (1) When comparing State level statistics from a census for two groups of States, the statistics are net figures which show the effect of several factors including migration, differential birth rates, differential missed rates, etc. Using these data to detect a coverage difference between States for a particular reason would require that reason to have a very large effect on the net figure to compensate for other effects. (2) It is very difficult to obtain information on the precise requirements of the welfare programs of the various States, and even more difficult to obtain information for 1960. The rules are constantly changing and there is no indication of how long it takes for new rules to go into effect. (3) There is also no way to measure the level of enforcement in the States in 1960. (4) In all the lists the majority of "strict rule" States are in the South, and therefore, geographic differences may confuse the issue.

[1] See Appendix B for the data.

[2] The statistical test used to compare the data was the Mann-Whitney U Test in which the alternate hypothesis is that States without "strict" welfare rules have higher sex ratios. (See Siegel, Nonparametric Statistics.)

[1] Results of the Mann-Whitney U Test showed that the Null hypothesis could not be rejected at .05 level. (See Appendix C for the data.)

Table 1.--MEDIAN SEX RATIOS FOR "STRICT RULE" AND "NONSTRICT RULE" STATES BY THE ALTERNATIVE LISTS OF STATES, TOTAL UNITED STATES

Selected age-color groups	List 1	List 2	List 3	List 4	List 5
Total U. S., Nonwhite, 20-49, State Level Data, 1960 Census					
"Strict rule" States	87.6	86.8	86.8	85.9	85.9
"Nonstrict rule" States	101.1	101.1	101.1	101.6	101.6
Total U. S., Negro, 20-49, State Level Data, 1910 Census					
"Strict rule" States	101.7	91.7	93.4	92.6	94.2
"Nonstrict rule" States	112.8	113.6	113.6	113.6	113.6

Table 2.--PROBABILITIES THAT "STRICT RULE" AND "NONSTRICT RULE" STATES DO NOT DIFFER IN SEX RATIOS, BY ALTERNATIVE LISTS OF STATES, TOTAL UNITED STATES

Sex ratios for selected ages	List 1	List 2	List 3	List 4	List 5
Total U. S., Nonwhite, 20-49, State Level Data, 1960 Census	.0003	.0022	.0005	.0009	.00016
Total U. S., Negro 20-49, State Level Data, 1910 Census	.0116	.0010	.0005	.0038	.0016

Table 3.--MEDIAN SEX RATIOS FOR "STRICT RULE" AND "NONSTRICT RULE" STATES BY ALTERNATIVE LISTS OF STATES, SOUTH REGION

Selected age-color groups	List 1	List 2	List 3	List 4	List 5
South Region States, Nonwhite, 20-49 State Level Data, 1960 Census					
"Strict rule" States	85.2	84.9	84.9	84.9	84.9
"Nonstrict rule" States	90.8	90.8	90.8	93.6	93.6
South Region States, Negro, 20-49, State Level Data, 1910 Census					
"Strict rule" States	94.2	91.3	93.4	91.3	93.4
"Nonstrict rule" States	100.6	100.6	100.6	103.8	103.8

Table 4.--PROBABILITIES THAT "STRICT" AND "NONSTRICT" STATES DO NOT DIFFER IN SEX RATIOS, BY ALTERNATIVE LISTS OF STATES, SOUTH REGION

Sex ratios for selected ages	List 1	List 2	List 3	List 4	List 5
South Region States, Nonwhite, 20-49, State Level Data, 1960 Census	a/	.268	.217	.071	.044 b/
South Region States, Negro, 20-49, State Level Data, 1910 Census	a/	.026 b/	.064	.143	.133

a/ Exact probabilities not available in tabulated data. Results of test indicate no significant difference at the .05 level.

b/ Significant at the .05 level.

Table 5.--MEDIAN SEX RATIOS FOR "STRICT RULE" AND "NONSTRICT RULE" STATES BY THE ALTERNATIVE LISTS OF STATES, CENTRAL CITIES OF SMSA'S FOR TOTAL U. S. AND SOUTH REGION

Selected age-color groups	List 1	List 2	List 3	List 4	List 5
Total U. S., Nonwhite, 20-39, Central Cities of SMSA's by State, 1960 Census					
"Strict rule" States	81.5	81.5	81.5	79.4	79.4
"Nonstrict rule" States	87.6	86.7	86.7	88.6	88.6
South Region States, Nonwhite, 20-39, Central Cities of SMSA's by State, 1960 Census					
"Strict rule" States	79.0	79.2	79.2	79.2	79.2
"Nonstrict rule" States	80.1	80.1	80.1	81.6	81.6

Table 6.--PROBABILITIES THAT "STRICT" AND "NONSTRICT" STATES DO NOT DIFFER IN SEX RATIOS, BY ALTERNATIVE LISTS OF STATES, CENTRAL CITIES OF SMSA'S

Sex ratios for selected ages	List 1	List 2	List 3	List 4	List 5
Total U. S., Nonwhite, 20-39, Central Cities of SMSA's by State, 1960 Census	.0018	.0735	.0071	.1271	.0026
South Regions States, Nonwhite 20-39, Central Cities of SMSA's by State, 1960 Census	a/	.396	.362	.321	.356

a/ Exact probabilities not available in tabulated data. Results of test indicate no significant difference at the .05 level.

APPENDIX A

Composition of Lists of "Strict Rule" and "Nonstrict Rule" States

	LIST 1			LIST 2			LIST 3			LIST 4			LIST 5		
	yes[1]	no[2]	unknown[3]	yes[1]	no[2]	unknown[3]	yes[1]	no[2]	unknown[3]	yes[1]	no[2]	unknown[3]	yes[1]	no[2]	unknown[3]
SOUTH															
Alabama	x			x			x			x			x		
Arkansas	x					x	x					x	x		
Delaware		x			x			x			x			x	
D. C.	x			x			x			x			x		
Florida	x			x			x			x			x		
Georgia	x			x			x			x			x		
Kentucky		x			x			x			x			x	
Louisiana	x				x			x				x			x
Maryland			x		x			x				x			x
Mississippi	x			x			x			x			x		
North Carolina	x					x			x			x			x
Oklahoma	x					x			x			x			x
South Carolina	x			x			x			x			x		
Tennessee		x				x			x			x			x
Texas	x					x	x					x	x		
Virginia	x					x			x			x			x
West Virginia	x				x			x				x			x
OTHERS															
Alaska[4]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Arizona	x					x			x			x			x
California		x			x			x			x			x	
Colorado		x			x			x			x			x	
Connecticut		x		x			x					x			x
Hawaii[4]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Idaho		x			x			x			x			x	
Illinois		x			x			x			x			x	
Indiana		x			x			x			x			x	
Iowa	x				x			x				x			x
Kansas		x			x			x			x			x	
Maine		x			x			x			x			x	
Massachusetts		x			x			x			x			x	
Michigan	x			x			x			x			x		
Minnesota		x			x			x			x			x	
Missouri	x					x			x			x			x
Montana		x			x			x			x			x	
Nebraska		x			x			x			x			x	
Nevada	x				x			x				x			x
New Hampshire		x				x			x			x			x
New Jersey		x			x			x			x			x	
New Mexico		x				x			x			x			x
New York		x			x			x			x			x	
North Dakota		x			x			x			x			x	
Ohio		x			x			x			x			x	
Oregon	x				x			x				x			x
Pennsylvania		x			x			x			x			x	
Rhode Island		x			x			x			x			x	
South Dakota		x			x			x			x			x	
Utah		x			x			x			x			x	
Vermont		x			x			x			x			x	
Washington		x			x			x			x			x	
Wisconsin		x			x			x			x			x	
Wyoming		x			x			x			x			x	

[1] "Strict rule" state

[2] "Nonstrict rule" state

[3] Not identified

[4] Omitted from consideration

APPENDIX B--SEX RATIOS FOR THE NONWHITE POPULATION, 1960 CENSUS AND FOR THE NEGRO POPULATION, 1910 CENSUS, AGES 20-49 BY STATE

	Sex Ratio (1960)	Sex Ratio (1910)
	Nonwhite, ages 20-49	**Negro**, ages 20-49
SOUTH		
Alabama	81.12	92.56
Arkansas	79.96	101.68
Delaware	96.50	106.88
District of Columbia	87.62	79.61
Florida	94.51	118.66
Georgia	83.91	90.06
Kentucky	90.79	100.64
Louisiana	84.04	97.95
Maryland	94.05	97.46
Mississippi	80.08	94.17
North Carolina	89.41	83.89
Oklahoma	85.23	112.56
South Carolina	85.90	86.24
Tennessee	82.45	89.47
Texas	88.09	98.49
Virginia	96.86	93.60
West Virginia	76.76	156.40
Weighted Average for South	87.09	94.42
OTHERS		
Alaska[1]	-----	-----
Arizona	98.32	104.92
California	101.08	114.68
Colorado	102.73	106.69
Connecticut	93.53	90.89
Hawaii[1]	-----	-----
Idaho	116.92	179.49
Illinois	88.11	113.60
Indiana	89.66	107.07
Iowa	92.25	126.31
Kansas	101.72	110.79
Maine	153.89	115.36
Massachusetts	99.70	100.33
Michigan	91.43	113.89
Minnesota	104.19	161.88
Missouri	83.64	108.65
Montana	104.47	137.47
Nebraska	99.84	130.88
Nevada	105.13	105.202
New Hampshire	150.84	104.48
New Jersey	89.20	94.21
New Mexico	97.87	130.13
New York	82.16	92.98
North Dakota	101.69	193.33
Ohio	90.05	112.78
Oregon	104.85	172.54
Pennsylvania	85.86	100.20
Rhode Island	108.93	100.44
South Dakota	102.04	149.49
Utah	112.05	166.09
Vermont	133.80	383.76
Washington	112.98	184.37
Wisconsin	97.32	105.204
Wyoming	101.60	283.44
Weighted Average for U. S.	88.68	96.01

[1] Omitted from consideration.

APPENDIX C--PROPORTION OF NONWHITE HOUSEHOLDS (IN RENTER-OCCUPIED UNITS) HEADED BY FEMALES FOR CENTRAL CITIES BY STRICTNESS OF WELFARE RULES: 1960[1]

Central cities of SMSA's by strictness of welfare rules[2]	2-or-more-person households (1)	Total female head (2)	Total female head income less than $3,000 (3)	$\frac{(2)}{(1)}$	$\frac{(3)}{(1)}$
"Strict rule" states					
Birmingham, Alabama	18,681	5,005	4,394	.268	.235
Washington, D. C.	60,225	15,132	8,905	.251	.148
Atlanta, Georgia	28,929	8,388	6,320	.290	.218
New Orleans, Louisiana	37,785	10,670	8,977	.282	.238
Detroit, Michigan	65,552	17,554	13,567	.268	.207
Kansas City, Missouri	9,412	2,911	2,274	.309	.242
St. Louis, Missouri	34,574	10,750	8,319	.311	.241
Dallas, Texas	18,747	4,808	3,886	.256	.207
Houston, Texas	28,293	7,095	5,796	.251	.205
"Nonstrict rule" states					
Los Angeles, California	60,921	16,119	10,342	.265	.170
Oakland, California	12,227	3,612	2,752	.295	.225
San Francisco, California	21,012	4,663	2,871	.222	.137
Chicago, Illinois	155,744	41,304	27,546	.265	.177
Indianapolis, Indiana	11,944	2,859	2,032	.239	.170
Newark, New Jersey	28,335	7,363	4,378	.260	.155
New York, New York	234,871	69,591	40,002	.296	.170
Cincinnati, Ohio	19,676	5,521	4,248	.281	.216
Cleveland, Ohio	40,346	9,727	7,033	.241	.174
Philadelphia, Pennsylvania	64,388	21,539	15,391	.335	.239
Pittsburgh, Pennsylvania	16,199	4,461	3,603	.275	.222
Memphis, Tennessee	25,800	6,698	5,840	.260	.226
Not identified					
Baltimore, Maryland	44,590	12,480	9,108	.280	.204
Honolulu, Hawaii	23,538	3,138	1,613	.133	.069

[1] Cities with less than 25,000 units with nonwhite household heads not reported.

[2] "Strict" or "nonstrict" states according to list 1.

SOME STATISTICAL ASPECTS OF EDUCATIONAL EVALUATION

Charles Schotta, Virginia Polytechnic Institute
Richard B. Hoffman, State University of New York at Buffalo

Introduction

This paper discusses some statistical considerations underlying educational evaluation. We first point out the objectives of educational evaluation. Then we evaluate the existing set of procedures for producing such estimates from the standpoint of classical probability theory. Next we present some empirical evidence in support of our criticism of existing methods of securing educational evaluation. Some major alternatives to existing evaluation practices are then discussed. We conclude by exploring the essentially Bayesian nature of educational evaluation practices and by delineating some important directions for future research efforts.

I. The Objectives of Educational Evaluation Systems

Typically an educational evaluation, the "grade", is an attempt to measure or to identify the result of a course of instruction. Additionally this measure is used to discriminate between students for a number of purposes.[1] It appears that the manner of securing this educational evaluation can be described as follows: during a course of instruction of some time-interval, say a semester of 18 weeks, the process of instruction (which we, as economists, interpret as the production of human capital) is interrupted several times to administer a test instrument.[2] We will discuss the interpretation of the test instrument in our view of educational evaluation in a later section of this paper. The numerical (or alphabetic code) result of each instrument is weighted by some explicit or implicit function to produce a point estimate. This estimate is taken to represent the measure of the quality of human capital created in the course of instruction, an identification of the product itself, a specification of the intensity of the student input, or various combinations of these. The exact form of the relationship between these and the "grade" does not appear to be a simple one.

This description appears to be an accurate one regardless of the type of test instruments administered. To sum up, the essential features are (1) a rather large population of "learning" or "teaching" or "production" periods, (2) a fixed, usually small (relative to the above) number of "test points", (3) some weighted mean of the value of the "test points" is represented as a meaningful measure of the value of the units in (1), (4) the "grade" produced as (3) is used to rank whatever it is the course of instruction produced, that is, we differentiate the human capital that the student ends up with by using the grade set produced for a class of students.

The striking feature of this procedure is that, on the basis of a sampling technique of this kind, inferences about differences in human capital, with respect to quality and quantity, are made and rather precise decisions are based on these inferences. The usual consideration of variance of the estimates is totally lacking. Here an example might be useful. Suppose we have an introductory statistics course in which we hold 45 lectures during the semester. We sample the work with some instrument three times during the semester. We secure a numerical average for each student. Using some conversion scheme we translate this set of numerical averages into relative alphabetic-coded rankings. Invariably one must make a decision whether to assign a "B" code to the 79 and a "C" code to the 78 or to define the code class demarcation point one, two, or some number of units lower or higher. Once, however, the "grade" has been assigned, it is clearly inferred that you have "different" commodities.

We believe that one can properly describe the objective of educational evaluation as the securing of a ranking of the quality and quantity of the creation of human capital during a course of instruction which will serve as a proxy for the measurement of the efficiency of output of this capital "production" process. Regardless of the initial intent of evaluators, the result of their evaluation is typically used in resource allocation decisions.

II. A Sampling View of Educational Evaluation

Suppose we have the "grade" determined as we suggest it presently is determined. This is a point estimate with no variance estimate. Let us represent a course where test instruments x_1, x_2, x_3 are given and the explicit weighting function is .25, .25, .5. These test instrument measures might be thought of as identifying the state of each student at, say, points y_{12}, y_{36}, and y_{45} for a course of instruction, Y with 45 time points.

$$(1)\quad E = .25\, x_1 + .25\, x_2 + .5\, x_3$$

and is the estimate of the mean, y, of the actual value of points y_1 through y_{45}. E then has, although almost no educational evaluation procedures consider it, a variance estimate.

If one calculates this variance and expresses it as the standard deviation of E, then, based on the sample size, one can, at various probability levels, calculate confidence intervals around E for each student.

Suppose for the i^{th} student we have

$$(2)\quad E_i \text{ with } \sigma_E^i$$

and we wish to form the confidence interval at the .95 probability level. In the above example, as is usual in educational evaluation, the sample (the number of test instruments) is less than 10 per cent of the population (the number of production points) hence no population correction factor needs be applied.[3]

Let us represent the true measure of the human capital creation which occurs in the 45 units as G_i.

Then, (3) $P[E_i - t\sigma_x^i / n^{1/2} < G_i < t\sigma_x^i / n^{1/2} + E_i] = 1 - \alpha$

For the case in question for the .95 probability level this would be

(4) $P[E_i - 1.96\sigma_x^i / n^{1/2} < G_i < 1.96\sigma_x^i / n^{1/2}] = .95.$

Although in a later section of this paper we present some summary results of viewing an educational evaluation in this way, we present below a hypothetical numerical example to illustrate our point. Assume: A = 90 - 100, B = 80-90, C = 70-80, n = 3, at the $1 - \alpha = .95$, the grades would be, for the given means,

	Mean Score	SD	Grade Rank
Student 1	75	4	C
Student 2	78	4	C or B
Student 3	85	7	A, B, or C

These examples illustrate that when one views the testing process as a sampling process, when inferences are required at a specified level of probability, it is often quite unclear that class rankings are unequivocal. Throughout our analysis we have assumed that the numerical measure attained on a given test instrument represents a zero variance point estimate of the state of the mind being tested.[4] This represents the most unfavorable assumption which can be made for the consequences of non-zero variance between test instrument scores upon which our criticism of current evaluation methods rests. If one relaxes this assumption to more closely reflect reality, then our criticism is broadened to include the necessity for probabilistic evaluation of the numerical measures for each test instrument.

From the above examples it can be seen that there exists a situation where rankings have questionable meanings and alphabetic codings appear to be representations of rankings which are themselves of doubtful value in further decisions.

III. Empirical Evidence on the Validity of Some Grade Rankings

The following section describes the results of applying the foregoing statistical analysis to several large section social science courses where three instruments were applied during a semester. Two different weighting functions were applied, one consisting of a pattern of .25, .25, and .5; the other an equal weight system. Both sets of instruments consisted of two 50 (MC) multiple choice item instruments and one 120 (MC) item instrument. The instruments were not applied at a random time during the course of instruction. It is assumed that the numerical score for each student for each instrument is an estimate of the state of the human capital created at the time of the application of the instrument. It is further assumed that the numerical score on each instrument has an associated variance of zero.

When evaluated in the fashion suggested in Secion III, out of a total number of respondents of 560, in four classes of size 80, 180, 145, 155, only 20 per cent of the students had confidence intervals at the .95 probability level which were contained entirely within the pre-determined numerical intervals for conversion to alphabetic grades.

When the students were grouped into sets according to whether their numerical averages and the associated confidence band (1) overlapped a grade interval higher than that in which their mean lay, (2) overlapped a grade interval lower than that in which their mean lay, and (3) overlapped both as in both (1) and (2) above, the distribution between the three classes was

	(1)	(2)	(3)	no overlap
number	203	156	89	112
Percent	36	28	16	20

In the test set of students, it was far more likely that either a grade too high or a grade too low was assigned than was the case that there was a probability that any one of three grades was probable.

Indeed, the cases (1) and (2) are the interesting problems since in case (3) it was, on the average, only necessary to lower the probability level to approximately .7 to shrink the confidence interval to lie either entirely within the grade interval or to join cases (1) or (2).

On the other hand, it was necessary to lower the probability level to .4 to eliminate overlapping in case (1), while it was necessary to lower the probability level to .5 to eliminate overlapping in case (2).

It should be stressed that even then, the problem of assigning different grades to individuals who, using some test for difference between two estimates of means of different populations, each with a variance estimate, do not appear to have different means, remains and is a major obstacle to a clear-cut interpretation of the rankings which result from alphabetic grade-code assignment.[5]

It was on the basis of the empirical evidence cited above that we came to our considerable skepticism concerning the quality of educational evaluations produced by the system of educational evaluation outlined in Section I which we take to be wide-spread,

IV. Some Alternatives to Present Educational Evaluation Systems

In this section we wish to consider some major alternatives to the present system of educational evaluation and to evaluate these proposed alternatives against the simple framework of sampling analysis in which the present system was presented. While some of the alternatives are in reality modifications to the present system to circumvent difficulties we have pointed out above, others represent radical proposals for reform of educational evaluation.

A simple modification of present practice which would greatly reduce the variance of the numerical grade estimate is an increase in the number of instruments applied during a specific course of instruction. If this is also coupled with random selection of the specific points of sampling one can more easily reconcile the interpretation of the grade with practice in statistical quality control.

It must be clearly recognized that the assumption that the items on an instrument produce an estimate of the state of the population being sampled with a zero variance is unlikely to be true in practice. Hence the overall quality of the measure of what occurs in a course of instruction is affected not only by the number of instruments and the manner of their application and weighting, but also by the number of items per instrument and the variance associated with the "score" on each instrument. Unfortunately, although we can estimate the sample variance for the results of several instruments,we cannot do so for the items on an individual instrument. Many educational psychologists choose to regard the numerical measure of the items on an instrument as data without observation errors. If this is the approach selected then this is tantamount to accepting the zero variance nature of an instrument "score". The quality of the educational evaluation for a course of instruction is then, from the standpoint of the approach we are taking toward present procedure, independent of the number of items on a given instrument and is determined only by the number of instruments and the variance particular to each respondent being evaluated.

The suggestion that the number of instruments applied be increased leads to an interesting conclusion. Suppose that the application of an instrument is analogous to "destructive testing" in quality control in the sense that while testing the production of human capital does not proceed. There is then a trade-off between additional accuracy and capital building. If we reduce each instrument to one item and and structure the course of instruction in such a way that after each information bit is presented an item is presented, we have programmed instruction. One completes such a course of instruction by a time pattern of binary conditions which eventually leads to the last information bit and item in the sequence. Clearly the results of such a procedure must be evaluated in some manner additional to the items which follow each information bit. Usually the procedure is similar to the standard system we describe with formal instruments consisting of a large number of items being applied several times during the course. In principle, then, there is no difference between the result of such a procedure and that from a conventional course. Occasionally a "grade" for PI course will be derived from a measure of the percentage of successes on individual items. Although this procedure has some interesting aspects, we do not examine it.

Another major alternative is the system where only one instrument is applied for a course of instruction or for a specified sequence of courses. Such measures cannot be used for inferences about the capital creation process but only infer something about a state of the human capital created and in existence at the end point. If we take the view of the human capital which has been created and the physical equipment as being analogous to software, data, and hardware, then we are sampling the contents of bits in core, the ability of a program to call out the correct subroutines, and the logical structure of the circuitry. We think this is a reasonable view to take of such an examination procedure since the most usual objective given for the "comprehensive exam" is that "we will find out what he knows and doesn't know."

V. Educational Evaluation as a Bayesian Decision

There is another interpretation of the educational evaluation process as it is practiced today which might be related to Bayesian decision theory. Suppose we regard the "course average" as the most probable number describing whatever we wish to measure rank for a student in a course of instruction. Given the numerical interval for alphabetic coding, if a student score lies in the middle of the grade band, we are likely to answer the implicit question "what is the probability that his 'true' measure (if one exists) is high enough (or low enough) to give him the higher (or lower) grade" by saying "very low" and to regard this student as a "solid C" or whatever grade is in question. When, however, we have a student whose "course average" is on or near one or another end of the grade band, the a priori probability that his grade could be the higher or lower grade is considerably increased. For the student on the endpoint of the grade band, we might even feel that either grade could be correct. We seek additional information to make the decision. Almost invariably we do not replicate the experiment which produced the score under consideration. Rather we look at the pattern of scores on the instruments to see if it is "rising", "falling" or some such thing. We try to think about the personality of the student. We consider "special" factors[6] and make a decision. The empirical evidence concerning the degree of overlap of the grade interval and the confidence interval presented in Section III suggests that this type of decision process is the most prevelant manner of resolving much difficulties discussed above lends support to the interpretation of the grading process as a Bayesian decision.

V. Research Areas

We would like to mention what we see as some areas for further research into the question of educational evaluation. We see this as important simply because of the resourse allocation decisions which are made on the basis of educational evaluations. Resource misallocation will result to the extent these evaluations are erroneous.

First, underlying the realiability of evaluations is the reliability of instruments. We are forced to regard the "score" for an instrument as zero-variance estimate or as data. The problem of the reliability of items and of instrument construction are not within our province but must be recognized as fundamental to successful evaluation.

It must, however, be recognized that the whole area of the design of the sampling plan, the selection of the weighting function, and the selection of the method of securing aggregation of course grade measures into larger measures is crucial to the production of quality measures for use in resource allocation[7]

---

1. Based on the relative ranking in a class, a conversion to some conventional grading system such as letter grades will be made. These letter grades will be subjected to additional transformations and operations and the results will be used for such purposes as deciding whether fellowships and scholarships will be given, whether or not a person will be continued in the educational process, what position to assign to a person in the job structure, whether a given male will be drafted or not, etc.

2. For a discussion of education as the process of "creating" human capital see T. W. Schultz, "Investment in Human Capital," Am. Econ. Rev. (March, 1961).

3. One must recognize here that number of items per test instrument may be substituted to an extremely limited degree for additional test instruments. However, since the number of items on an instrument must necessarily be extremely small relative to the population of information to be sampled and the items are not likely to be independent, the effective sample size is not likely to be markedly affected by an increase in the number of items per test instrument.

4. Essentially the score for an instrument represents the weighted sum and remainder of a set of binary values.

5. These results were obtained as a by product of computer programs written to implement weighting functions and alphabetic grade-code conversions. It was merely necessary to add variance calculations and confidence interval calculations and grouping operations to produce the above results.

6. It should be noted that considering such factors for one student violates what we call the principal of "horizontal equity". This term, widely used in public finance theory in connection with tax loads, simply means that one must treat equals alike. As it is applied in taxation much attention is given to determining classes of equals so that they can be taxed alike. In connection with educational evaluation we interpret its application to mean that unless information of the same type is available and considered for every student with the same weight, it should not be considered for any. For example, consider the student who "blows" the final exam, then comes to you with the news the next day that a relative died or that he had a case of the 24-hour flu prior to the exam. To adjust his grade by assigning a positive weight to the new information would violate the principle of horizontal equity unless you accumulated information on the state of all other students who are included in the ranking with respect to these two conditions. Rigid adherence to this principle would, of course, have the very undesirable effect of preventing the evaluator from considering information of relevance for a decision.

7. For the results of a simulation model of the educational process which is designed to measure the impact of alternative sampling plans, weighting functions, and aggregation procedures on the final amount of human capital created by ascertaining the effect on decisions on who remains in the system and who is ejected from the system, see C. J. Goetz and C. Schotta, Jr., "Quality Control in the Production of Human Capital: A Simulation Study," paper to be presented at the Operations Research Society of America meetings, Philadelphia, November 4, 1968. This model ascertains the results of items in such a way that exam scores are approximately zero variance data since the "value" of each simulated "mind" in the simulated "population" can be ascertained and compared with the "value" derived from the instrument. The simulation study is based on the central theme of our earlier paper, Schotta and Hoffman, "A Priori Decision Functions for Education Evaluation" presented at the Operations Research Society of America meetings, New York, May 31, 1967.

THE TREND OF OBESITY IN THE USA

by

P. V. Sukhatme
Food and Agriculture Organisation, Rome

1. INTRODUCTION

The importance of controlling obesity for the health of man in the middle and later age groups is now fairly well recognised. While hereditary factors may contribute to the incidence of obesity in the population, overeating and under exercise are also among the important factors influencing it. The purpose of this paper is to study the incidence of obesity and its likely trend in the context of growing mechanization, keeping in mind the trends in food supply and population growth.

2. OVERNUTRITION

In an affluent society food supply exceeds very considerably the food needed to maintain a healthy active life. This is demonstrated by the data in Table 1. The Table shows that as against a calorie requirement of 2600 per caput per day, the calorie supply available during 1954/56 was of the order of 3200, thus exceeding requirement by about 20%. The Table also shows that during 1964/66 the calorie supply remained more or less the same as in 1954/56, but owing to a downward revision of calorie requirement in 1964 consequent on the changed pattern of activity, the excess of supply over requirement further increased, being 33%. This excess is seen to occur in all income groups (vide Table 2) and shows that there is no insufficiency of calories in the USA, even in the poorest classes.

Much of the excess supply must clearly be ascribed to waste which appears to be more than the 10% allowed for in estimating requirement at retail level from that at the physiological level. But there can be no denying that an excess supply as large as that shown in Tables 1 and 2 must favour overeating. This is amply borne out by the data on food consumption collected during the nationwide household surveys in 1955 and 1965 and presented in Table 3. As the Table shows, in 1955 three out of every four households had a calorie supply per nutrition unit exceeding 3500. This proportion was even higher in the year 1965 being four out of five. This does not however mean that the population covered by three out of every four households in 1955 or four out of every five households in 1965 can be regarded as overnourished any more than the population covered by one out of every four in 1955 or one out of every five households in 1965, which is found to fall short of the corresponding requirement, can be considered to be undernourished. Clearly some people will need less than the stipulated average calorie requirement while others may need more depending upon the extent to which the different individual factors, including physical activity, deviate from those of the 'reference' man. The only way of estimating the incidence of overnourished in the population is to evaluate o given by

$$o = \int_{\frac{x}{y}>1} f\left(\frac{x}{y}\right) d\left(\frac{x}{y}\right)$$

where x represents the calorie intake, y the corresponding requirement and $f(\frac{x}{y})$ the distribution function of calorie intake relative to requirement of individuals in the population. Unfortunately, there is no information available on the distribution function $f(\frac{x}{y})$ for individuals. The available nationwide data collected during 1955 and 1965 related to intake of households per nutrition unit and not individuals. We can however approximate to the expression o by evaluating o' given by

$$o = \int_{\bar{x} \geqslant c - 3\sigma_{\bar{y}}} g(\bar{x})\, d(\bar{x})$$

where $\bar{x}$ = calorie intake of the households per nutrition unit

$\bar{y}$ = calorie requirement of the household per nutrition unit*

c = calorie requirement of the 'reference' man

and $\sigma_{\bar{y}}$ = standard deviation of $\bar{y}$

Available data show that the standard deviation of energy expenditure among healthy active adults of the reference type is roughly 500. Since the average size of the household is approximately 2.5 nutrition units, we may conclude that $\sigma_{\bar{y}}$ is roughly 300.

Clearly in a well-fed society we would have expected most households in 1955 to have had calorie supplies per nutrition unit between two limits 3500 ± three times 300, that is 2600 and 4400. If we use the revised recommendations of the Food and Nutrition Board (1964) this range would be 2300 to 4100.

In actual fact, as Table 5 shows, a large proportion of the households is seen to have a calorie intake exceeding the limits given above. It would appear that the proportion of overnourished approached 50% in 1955 and if anything has further increased in 1965 to 57%. It is of course possible that the calorie supply available for consumption has been overestimated since much more food is probably wasted than has been allowed for. It is also possible that the requirement is underestimated in the sense that the average of energy expenditure on the physical activities of a person does not correspond to the level of moderate activity visualized for the reference man. Even with these errors the evidence appears conclusive that nearly half of the population in the USA, is overeating and that, if anything, the incidence of overnutrition in the population is increasing.

3. TRENDS OF ENERGY EXPENDITURE RELATIVE TO INTAKE

That food supply exceeds needs is easily verified by the many obese people we see everywhere in the advanced countries. Available evidence shows that the obese on average eat no more than non-obese persons of comparable age and occupation. It follows that obese people must be physically less active and spend less energy than the non-obese. If then we can ascertain the rate of energy expenditure and its trend during daily occupations we should be able to infer the trend of obesity as well. The best way is to study the data of surveys of physical activity of working populations, but such surveys are difficult to organize (Passmore 1962). In the absence of data for such surveys we can only attempt an indirect and rough assessment of the extent of reduction of energy expenditure by analysing the shift in the pattern of occupations and reduction in working hours of the working population for countries for which such data are available.

A glance at Table 4 on the classification of employed persons by occupation shows that there is a significant shift in the USA in the pattern of occupation since 1950 (US Department of Commerce, 1966). Thus, whereas in 1950 white collar workers accounted for about 37% of the total employed persons, they now exceed 44% of the total, thereby suggesting that more people are now engaged in activities which require less energy than that needed by the average adult worker. The decrease over the same period in the proportion of people employed in occupations requiring relatively larger energy expenditures supports this trend, e.g., whereas over 18% of the working force in 1940 was employed on farm work, such workers currently form only 6% of the total employed. The shift shows that energy expenditure per adult of the working population must have decreased over the last 15 years. Using known rates of energy expenditure of physical activity in different occupations and regrouping them into white collar, blue collar, service and farm workers, as shown in Table 4, we estimate that the energy expenditure on work in daily occupations has decreased by about 0.5 calories per minute on 8-hour working time, or roughly by 200 to 300 calories per day. It is likely that over and above this reduction more power has become available to replace in part the manual work done in 1950. It is also likely that with industries and workshops already mechanized in 1950 a further reduction on this gain may not be significant in the case of the USA. On a conservative basis we may say that the daily energy expenditure of the adult population on work in employment appears to have decreased at the rate of roughly 15-20 calories per annum. The revision of the calorie requirement scale announced by the Nutrition Committee of the USA, reducing the requirement of the 'reference' man and woman by 300 and 200 calories respectively, accords with these findings (Food and Nutrition Board, 1964).

While food requirements have undoubtedly decreased, available evidence indicates that the trend and pattern of food supply available for consumption in the USA has also changed over the last 15 years (FAO, 1963, 1966). However, the reduction in calorie supply appears to be of a smaller order (Table 5). The Table shows that the daily calorie supply per caput has remained steady or has gone down only slightly. What we need, however, is not the trend in per caput food consumption but the trend in food consumption of the adult population employed in work. An approximate estimate of this latter trend can be obtained by adjusting the data for food consumption per caput of total population by allowing for the physiological needs of children and old people in accordance with the recommendations of the Calorie Requirements Committee. Studying the trends so derived and allowing for the margin of error inherent in the estimation of food supply data, we may conclude that calorie supply per adult worker has probably decreased by up to 100 calories over the last 15 years. We conclude that the reduction in daily calorie supply is of a smaller order relative to the reduction in energy expenditure and that the difference can be placed at between 10 and 15 calories per annum.

It would be hazardous for one who is not a physiologist to attempt to translate these findings in terms of the trend of gain in bodyweight, since I have no knowledge whether all this excess supply goes in the formation of bodyweight, whether any part of it is excreted and how the relationship between excess calorie intake and gain in bodyweight changes at different levels of bodyweight. I would merely add what looks to me simple arithmetic, that if one were to assume all this excess in calorie intake were to be laid down as fat, an adult on average would be gaining in bodyweight to the tune of two thirds to 1 lb. per annum. Over the 14 year period from 1950-64 this would be equivalent to a gain in bodyweight of 10 to 15 lbs. per male adult.

4. TRENDS IN BODYWEIGHT

Direct evidence of gain in bodyweight is provided by data from surveys of physical measurement of the adult population. Such data for the USA from surveys carried out during the period 1941-63 are set out in Tables 6 and 7. Table 6 sets out the trends in average weight derived from the data collected by the National Centre for Health Statistics (1965, 1966) and American College Health Association Research Committee, while Table 7 shows the trend based on the data collected by the Metropolitan Life

Insurance Company and 26 other insurance companies in the USA. An examination of these tables shows that age for age there has been an increase in weight of the order of between 1-1½ lbs. per adult per annum. Table 6 shows a bigger gain than that recorded in Table 7. It is likely however that the data recorded in Table 6 for 1948-50 by the American College Health Association Research Committee relate to strata of society which have higher educational and social attainments and for this reason is not wholly comparable with the cross section of the population covered in 1960-62. In all probability, the difference is an overestimate of the actual gains even though the measurements have been adjusted to a comparable basis. The data presented in Table 7 have the merit that they are collected on a comparable basis from all insured persons numbering several thousand but the trend may reflect in part the differences in socio-economic strata likely to be different from those in the general population. Even with these drawbacks however the data leave little doubt that apart from adults becoming heavier and heavier, age for age, they put on significant and marked gains in weight as they become older. The gain is particularly large in males between the age groups 18/24 - 25/34. Thereafter the gains steadily diminish. Women are relatively more successful in controlling their weight up to the age group 25-34, but thereafter they too record gains in bodyweight comparable to or even exceeding those recorded by men.

5. INCIDENCE OF OBESITY AND ITS TREND

There is no generally agreed definition of obesity. Some workers, notably Mayer and Seltzer, 1965, have suggested definition based on caliper determination of skin-fold thickness; others use the simple approach of excess weight over published standards. The former is probably of greater value in clinical work; the latter however appears to be the more generally accepted line of thinking on obesity. Confining our attention to the latter, available literature shows that a person is considered to be obese when his weight exceeds 20% of the average weight of all adults in the population. Data of the nationwide survey conducted by the National Health Centre of the USA during 1960/62 show that the average weight of all adult males is 168 lbs. and that of all adult females is 142 lbs. The limits for obesity on the current convention therefore are 168 + .20 x 168, or approximately 200 lbs. for males, and 142 + .20 x 142, or 170 lbs. for females.

Table 8 shows the mean and the standard deviation of the distribution of weights of adults as observed in the nationwide survey conducted in the USA during 1960-62. It also shows the proportion of adults exceeding the limit of 200 lbs. for males and 170 lbs. for females. It will be seen that 12% of the adult males have a weight exceeding 200 lbs. and 16% of adult females have a weight exceeding 170 lbs. In other words, by current criteria the incidence of obesity in the population in the USA is 12% among males and 16% among females. The Tables also show that the incidence of obesity increases with age up to 65. The increase in the incidence of obesity with age is relatively more rapid in the case of males than females.

Overweight and obesity are relative terms. To use the average bodyweight of all adults in determining the limits beyond which a person can be classified as obese seems to us to beg the question since the average weight itself is influenced by the proportion of obese persons in the population. What is needed in our view is a reference weight based on a concept similar to that used by the FAO Committee on Calorie Requirements in defining calorie needs. According to this Committee a 'reference' man is 25 years old, weighs 70 kg. lives in a mean annual temperature of 10°C. and lives a healthy active life of moderate activity spending about 8 hours in working activities which are neither heavy nor sedentary, 8 hours in non-occupational activities and 8 hours rest in bed. Observations show that he consumes around 3200 calories per day. The calorie requirements for adults in other age groups are determined in relation to the deviation of bodyweight and of the degree of physical activity from those assumed for the 'reference' man.

Clearly, much the same approach is indicated in defining obesity. Heavier as an adult of 18-24 of today may be than he was 10 or 20 years ago, he is our 'reference' point and it is only in relation to the morbidity and mortality of healthy adults in this group that we can judge the significance of overweight and obesity. Experience has shown that conditions governing mortality and morbidity in life remain optimal when the bodyweight remains at the level it was during the age group 18-24 after a normal development characteristic of healthy active children. Adopting then our 'reference' man as a healthy active adult of the youngest adult age group of 18-24 experiencing the least risk of morbidity and mortality, we may define obesity as falling outside the normal range of weight of adults of the reference type.

But not all adults of the age group 18-24 can be regarded as healthy active adults subject to the same low risk of mortality and morbidity. Experience shows that the rates of mortality and morbidity increase as weight increases even in the age group 18-24. Carried to its logical conclusion this means that there is an optimal figure for weight for the youngest adult associated with least mortality, which when exceeded, makes a man increasingly overweight and susceptible to greater morbidity and mortality. But it is not practicable to define an optimal weight in terms of a single point in the scale of weights. Where, as in defining calorie needs, there is "a continuous progression from health to disease with increase in bodyweight" the only way of defining overweight and obesity is relative to the distribution of weight of healthy active adults of the youngest group experiencing the least mortality, and this is what we have ventured to suggest in the reasoning above. The reasoning is closely akin to that which has led the Metropolitan Life Insurance Company to define what they call the best or desirable weights associated with the least mortality and to give a range of desirable weights for different heights (Metro. Life Ins. Co., 1960). A study of these desirable weights shows the mean desirable weight is lower than the average weight of the youngest age group, being approx. 145 lbs. for males and 123 lbs. for females with the range extending from 112-204 lbs. for the former and 92-173 for the latter. Assuming normal distribution this would imply that the standard deviation of the weight of the 'reference' male adult is roughly of the order of 13-14 lbs. and that of the 'reference' female is of a like order. In other words, the probability that a healthy active adult of the reference group will exceed the mean desirable weight + three times standard deviation of the weight of the reference adult will be less than .01. If he should exceed this limit we can consider him to be obese and subject to a higher risk of mortality than the 'reference' adult.

We would have no hesitation in accepting the mean desirable weights as given by the Metropolitan Life Insurance Company as defining the mean weight of adults of the 'reference' type but adult males of 18-24 age group in 1960-62 were heavier by about 4-5 lbs. on the average than at the time when the Metropolitan Life Insurance Company prepared their tables. We shall be on the safe side if we say that our 'reference' man has a weight of around 150 lbs. with a standard deviation of 13-14. Likewise, our 'reference' woman will have a weight of around 125 lbs. with a standard deviation of around 13-14. Adopting this as our frame of reference then we may say that for the USA a limit of 150 + 3 s.d. or 190 lbs. can be considered as the limit to classify obese men and that of 165 lbs. to classify obese women.

In proposing these limits to define obesity we must utter a word of caution. The limits will change with the time consequent on the change of body size and the changing balance between energy expenditure and food intake. They are no more static than are the calorie requirements which also depend on the type of life people like to lead and their bodyweight.

Using the above limits we have calculated the incidence of obesity for the different age groups from the data collected by the NHS. These are shown in Table 8. As will be seen, the proportion of obese persons among males is slightly higher than we had estimated earlier using the conventional definition and amounts to some 21% and the proportion of obese persons among females is also 21%. The attached chart shows the method of estimating the incidence of obesity as also its extent.

We have projected the incidence of obesity for the year 1980 on the assumption of continued trend of excess intake of food relative to energy expenditure and the continued trend of increase in height and weight resulting partly from improved nutrition and partly from a tendency to marry outside one community. We find, using the normal form of distribution for bodyweight and coefficient of variation observed in 1960 in the N.H. Survey, that the incidence of obese people among the population is likely to increase by about one sixth by 1980. These calculations leave little doubt that the problem of obesity in the USA is likely to assume increasing significance unless food consumption further declines or physical activity during leisure and recreation is increased to compensate for the decrease in energy expenditure during working hours.

It is tempting to assess the impact of the trend in the incidence of obesity on the demographic picture but this is a complex task. Besides, the relevant data are not available. By way of example however we have calculated the potential gain in longevity to be had from preventing overweight and we find that the gain may amount to as much as two to three years in an expectation of 27 years for a man 45 years old (vide Table 9). Even in 1980, when life expectation will have increased at all ages on current trends of mortality as projected by the UN the potential gain in longevity to be had from preventing overweight remains much the same. The effect of controlling overweight on the life expectation of the population as a whole is necessarily smaller. Thus the potential gain in the expectation of life if overweight were controlled works out to a little over half a year (vide Table 10). Should the incidence of obesity increase by one sixth, as seems likely on current trends, the potential gain in life expectation will correspondingly increase, as can be readily seen from the Table. Even more significant than the potential gain in longevity is the reduction in the incidence of morbidity and of restricted activity through the development of what are termed by Linder (1966) "positive elements of health" in the form of a very considerable increase in the number of days of healthy living but we do not have data to illustrate these gains.

6. ACKNOWLEDGMENTS

I should like to express my thanks to members of my staff, particularly Messrs. W. Schulte, P.I. Petricevic, R. Calderoni, G. Celestini, N. Martone and Miss M. Homer without whose assistance it would not have been possible to prepare this paper in time.

* Calorie requirement of the nutrition unit is that of the reference man being 3500 at retail level in 1955 and 3200 at retail level in 1965.

REFERENCES

1. Food and Agriculture Organization of the United Nations (FAO), (1963): Third World Food Survey

2. Food and Agriculture Organization of the United Nations (FAO), (1966): Production Yearbook, Vol. 20, Rome

3. Food and Nutrition Board, U.S.A., (1964): Dietary Allowances, Publication 1146

4. International Labour Organization (ILO), (1966): Yearbook of Labour Statistics, Geneva

5. Linder, F.C., (1966): "The Health of the American People", Scientific American, Vol. 214

6. Metropolitan Life Insurance Company, (1960): "Mortality among Overweight Men and Women", Stat. Bulletin, Vol. 41

7. National Centre for Health Statistics, (1965): Weight, Height and Selected Body Dimensions of Adults, U.S.A., 1960-62, Series 11, No. 8

8. National Centre for Health Statistics, (1966): Weight by Height and Age of Adults, U.S.A., 1960-62, Series 11, No. 14

9. Passmore, R., (1962): "Estimation of Food Requirements", Journal Royal Stat. Society, Vol. 125, Part 3, Series A

10. Seltzer, C.C., and Mayer, J., (1965): "A Simple Criterion of Obesity", Clinical Nutrition, Vol. 38, No. 2

11. Society of Actuaries, U.S.A., (1959): Build and Blood Pressure Study

12. Sukhatme, P.V. (1961): "The World's Hunger and Future Needs in Food Supplies", Journal Royal Stat. Society, Vol. 124, Part 4, Series A

13. United States Department of Agriculture, (1957): Household Food Consumption Survey, 1955, Report No. 6

14. United States Department of Agriculture, (1967): Household Food Consumption Survey, 1965

15. United States Department of Commerce, Bureau of the Census, (1966): Statistical Abstract of the United States

Table 1

CALORIE SUPPLY COMPARED WITH CALORIE REQUIREMENTS

(per caput per day at retail level)

	Calorie Supply		Calorie Requirement	Supply as % requirement
1954-56	3160 1/	3200 2/	2600 3/	121
1964-66	3140 1/	3210 2/	2400 4/	133

Source: 1/ Food Balance Sheet for USA

2/ Household Food Consumption Survey, USA, 1955, 1965

3/ Based on recommended Dietary Allowances, 1958, Food and Nutrition Board, USA

4/ Based on recommended Dietary Allowances, 1964, Food and Nutrition Board, USA

Table 2

CALORIE SUPPLY BY INCOME LEVEL IN USA, 1955, 1965

(per caput per day at retail level)

Household disposable income ($/year)	1955	1965
Under 3000	3210	3120
3000 - 4999	3160	3180
5000 - 7999	3210	3230
8000 - 9999	3200	3280
10000 and over	3220	3300
TOTAL	3200	3210

Source: US Department of Agriculture (1957; 1967) Household Food Consumption Surveys 1955 and 1965

Table 3

DISTRIBUTION OF HOUSEHOLDS BY CALORIE SUPPLIES PER NUTRITION UNIT PER DAY - USA

(at retail level)

Calories per nutrition unit per day	Percentage frequency 1955	1965
Under 3000	12	14
3000 - 4000	28	26
4000 - 5000	28	27
5000 - 6000	16	17
6000 and over	16	16
Average calorie supply per nutrition unit per day	4390	4320
Average calorie requirement per nutrition unit per day	3500 1/	3200 2/
% households with calorie supply per nutrition unit exceeding 4400 (i.e. 3500 + 3 s.d.)	49	
% households with calorie supply per nutrition unit exceeding 4100 (i.e. 3200 + s.d.)		57

Source: 1/ Recommended Dietary Allowances for USA, Food and Nutrition Board, 1958
2/ Recommended Dietary Allowances for USA, Food and Nutrition Board, 1964

Table 4

PERCENT DISTRIBUTION OF ECONOMICALLY ACTIVE POPULATION BY MAJOR OCCUPATIONAL GROUPS - USA

Occupational Group	1940	1950	1960	1964
White collar workers	32.8	37.5	43.0	44.2
Blue collar workers	36.4	39.2	36.3	36.3
Service workers	12.5	10.9	12.6	13.2
Farm workers	18.3	12.4	8.1	6.3
Average calorie expenditure per minute of work	3.03	3.02	2.63	2.53
Index of trend in average calorie expenditure	100	100	87	84

Source: Statistical Abstract for USA, 1951 and 1965

Table 5

TRENDS IN DAILY PER CAPUT FOOD SUPPLY IN THE USA

	Calories from				
	Animal proteins	Vegetable proteins	Fats	Carbohydrates	Total Calories
Prewar	220	130	1130	1780	3260
1948-50	260	110	1220	1580	3170
54-56	275	105	1260	1520	3160
57-59	280	100	1250	1480	3110
60-62	275	100	1265	1460	3100
63-65	280	100	1270	1490	3140

Source: Derived from Production Yearbook (1966), FAO, Rome

Table 6

AVERAGE WEIGHT BY AGE FOR ADULT MALES AND FEMALES

Age Group (years)	Actual Average Weight* (lbs) 1948-50		1960-62+	
	males	females	males	females
18 - 24	151	122	160	129
25 - 34	156	125	171	136
35 - 54	162	136	172	145

*Adjusted to partial clothing without shoes

+U.S. National Health Survey 1960-62

Table 7

TRENDS IN AVERAGE WEIGHT, 1941-63*

Age Group males	1941 Metropolitan	1935-53 26 Companies	1963 Metropolitan
	.... weight (lbs)		
18 - 24	149	147	157
25 - 34	155	157	164
35 - 44	159	161	167
45 - 54	160	162	166
55 - 64	159	162	163

*Adjusted to partial clothing without shoes

Source: Statistical Bulletin, Metro.Life.Ins.Co.,Vol.47, 1966

Table 8

INCIDENCE OF OBESE AMONG ADULTS IN USA

Age Group (years)	Mean weight $\bar{w}$ (lbs)		Standard deviation (lbs)		Proportion of males exceeding*		Proportion of females exceeding	
	males	females	males	females	$\bar{w}$ + .20$\bar{w}$	$\hat{\bar{w}}$ + 3 s.d.	$\bar{w}$ + .20$\bar{w}$	$\hat{\bar{w}}$ + 3 s.d. of desirable weight
18 - 24	160	129	26	30	.06	.12	.08	.11
25 - 34	171	136	28	34	.15	.24	.16	.19
35 - 44	172	144	26	30	.15	.24	.19	.24
45 - 54	172	146	27	30	.15	.25	.21	.26
55 - 64	166	153	27	30	.11	.19	.28	.34
65 - 74	160	147	27	26	.07	.13	.18	.24
75 - 79	150	139	26	28	.03	.06	.13	.17
TOTAL	168	142	27	29	.12	.21	.16	.21

*These are derived on the assumption that the observed distribution of weights **are normally distributed with mean and standard deviation given in columns 2 and 3.** The proportion corresponds closely with the actual observed proportion as can be verified from the Tables given in the Report No.8 of Series 11 published by the National Center for Health Statistics, U.S.A.

Table 9

GAINS IN LONGEVITY FOR THE OVERWEIGHTS IF OVERWEIGHT WERE CONTROLLED

Expectation of Life in U.S.A. Males - 1959 and 1980

(1)	(2)	(3)	(4)	(5)	(6)	(7)
		1 9 5 9			1 9 8 0	
Age	All males excluding overweights 10% and more	Overweights only 10% and more	Years of life gained	All males excluding overweights 10% and more	Overweights only 10% and more	Years of life gained
0	67.0	64.3	2.7	71.8	69.2	2.6
25	45.5	42.8	2.7	48.7	46.1	2.6
45	27.4	24.9	2.5	29.7	27.3	2.4
65	13.0	11.3	1.7	14.2	12.4	1.6
% of population	79	21		75	25	

Source: Derived from data on mortality rates for USA and proportion of an excess mortality among overweights as reported in the Build and Blood Pressure Study, Society of Actuaries (1960).

Table 10

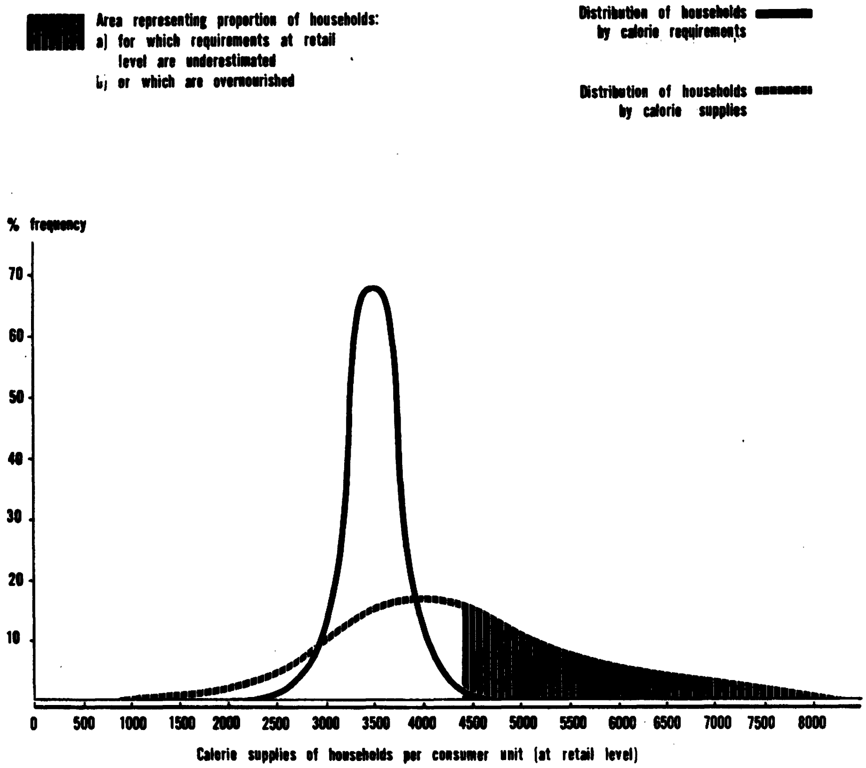
GAINS IN LONGEVITY FOR THE WHOLE POPULATION IF OVERWEIGHT WERE CONTROLLED

Expectation of Life in U.S.A. Males - 1959 and 1980

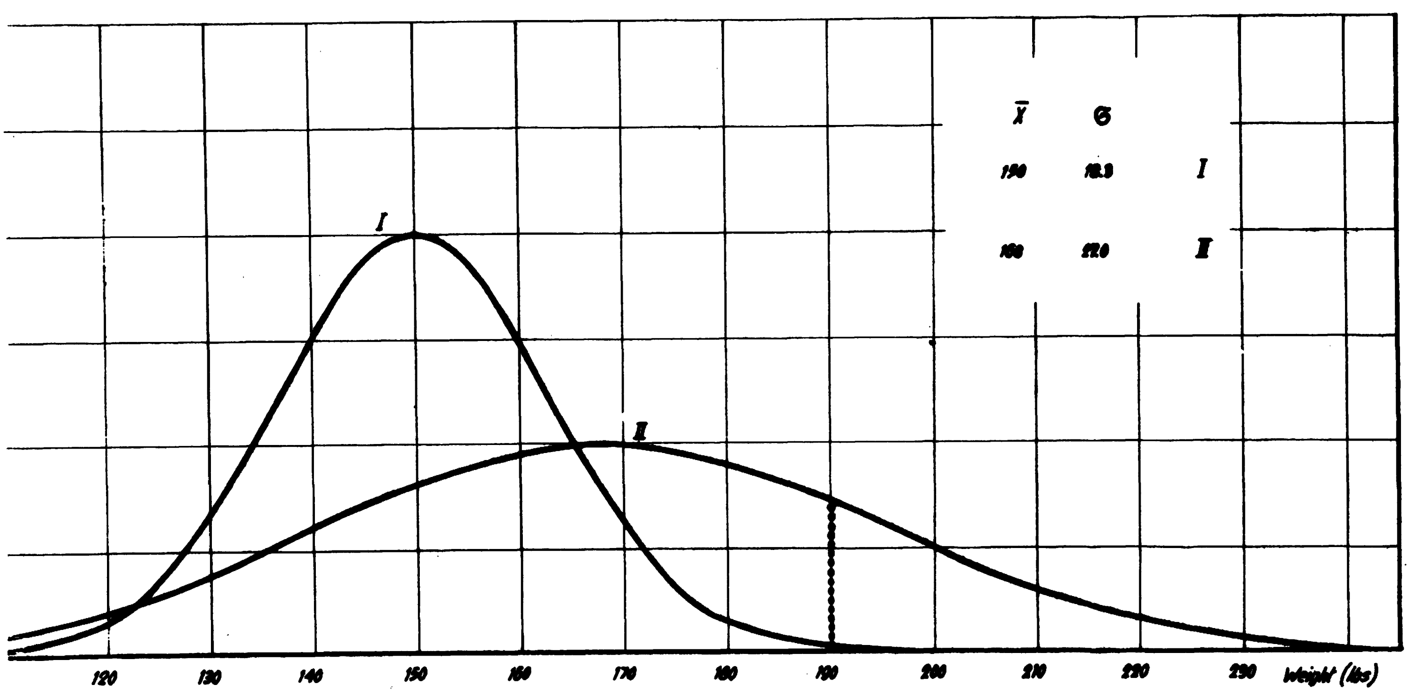
	1 9 5 9			1 9 8 0		
Age	All males	All males excluding overweights 10% and more	Years of life gained	All males	All males excluding overweights 10% and more	Years of life gained
0	66.4	67.0	0.6	71.1	71.8	0.7
25	44.9	45.5	0.6	48.0	48.7	0.7
45	26.8	27.4	0.6	29.1	29.7	0.6
65	12.6	13.0	0.4	13.7	14.2	0.5
% of population		79			79	

Source: Derived from data on mortality rates for USA and proportion of an excess mortality among overweights as reported in the Build and Blood Pressure Study, Society of Actuaries (1960).

CHART 2e - OVERNUTRITION IN THE U.S.A. (1955)



Incidence of obese among adult men - USA - 1960



XII

CONTRIBUTED PAPERS II

Chairman, ELIJAH L. WHITE, National Center for Health Statistics

INSTANT MATRIX DESIGN: STRAIGHT LINE WITH ERRORS IN BOTH VARIABLES

Irving H. Siegel, The W. E. Upjohn Institute for Employment Research*

My "supermatrix" and "normal-identity" approaches to the adjustment of linear data are applicable not only to the familiar cases in which the values of the dependent variable are subject to error but also to instances in which <u>all</u> the data may be inexact. At the 1967 ASA meeting, I illustrated the normal-identity approach to straight-line-fitting when both the observed y's and x's are uncertain (i.e., $y_i = Y_i - s_i$ and $x_i = X_i - t_i$). [1] I now present the supermatrix equivalent.

In the two adjustment procedures, residuals are introduced explicitly as unknown constants into the observation equations. These equations thus become observational identities of the form $y_i + s_i \equiv a + bx_i + bt_i$, or $y_i \equiv a + bx_i + \Delta_i$, where $\Delta_i = bt_i - s_i$ incorporates paired x and y residuals.

In the normal-identity approach, summary statements are first derived from the observational identities and then reduced to conventional "normal equations" by the suppression of certain aggregates. The eliminated aggregates correspond to plausible assumptions concerning the residuals. In the case considered here, there are three summary statements and four unknowns -- a, b, and two unsuppressed aggregates. A relation between the aggregates is accordingly postulated.

While the normal-identity approach features the "compression" of information, the supermatrix approach leaves the data intact. The observational identities and aggregate residual conditions are organized in unprocessed form (or nearly so) into a very large matrix system that is sufficient to determine a, b, and all the residuals. Although this supermatrix system is condensable to the usual normal equations, solution may be effected in any manner deemed expedient.

To derive the supermatrix system, we may start with what is left of the normal identities after the appropriate residual aggregates are equated to zero. As already noted, we have three equations in four unknowns, two of which are residual aggregates:

$$\left\{\begin{array}{rcl} \sum y &=& na + b\sum x \\ \sum xy &=& a\sum x + b\sum x^2 + b\sum tx \\ \sum y^2 + \sum sy &=& a\sum y + b\sum xy \end{array}\right\}$$

Next, we rewrite one of the unknown sums in terms of the other: $\sum sy = k\sum tx$, where k is a parameter. [2] This relationship is actually a disguised form of $\sum s^2 = k\sum t^2$, which states that the sum of squared deviations in the y direction is k times the sum in the x direction.

The first normal equation may also be written as $\sum \Delta = 0$, and the second and third yield $\sum (kx + by)\Delta = 0$. These two sums are aggregate residual conditions needed for completion of the supermatrix system, the design matrix of which is a square of the order $\overline{n+2} \times \overline{n+2}$.

The supermatrix system is this:

$$\left[\begin{array}{cc|cccc} \multicolumn{2}{c|}{} & 1 & 1 & & \cdots \\ \multicolumn{2}{c|}{0} & & & & \\ \multicolumn{2}{c|}{} & kx_1+by_1 & kx_2+by_2 & & \cdots \\ \hline 1 & x_1 & 1 & & & \\ 1 & x_2 & & 1 & & \\ \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & & & & \cdot \end{array}\right] \cdot \left[\begin{array}{c} a \\ b \\ \hline \Delta_1 \\ \Delta_2 \\ \cdot \\ \cdot \\ \cdot \end{array}\right] = \left[\begin{array}{c} 0 \\ 0 \\ \hline y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \end{array}\right]$$

The square design matrix is partitioned so that the packages of unprocessed information may easily be identified. If a numerical value is specified for k, a trivial amount of prior multiplication is required.

Since the design supermatrix contains b, solution for b yields a quadratic equation expressed in b! From this equation, shown in my paper of May 1967 and in some of the literature cited there, the value of b is readily ascertained.

---

*The author's views should not be ascribed to the Upjohn Institute.

[1] See I. H. Siegel, "From Identities to Normal Equations: An Easy Approach to Least Squares", <u>1967 Social Statistics Section Proceedings of the American Statistical Association</u>, pp. 354-356.

[2] In the 1967 paper (p. 355), k was inadvertently omitted.

PREDICTING DURATION SPECIFIC AVERAGES AND VARIANCES OF LIVE BIRTHS: APPLICATION OF A STOCHASTIC MODEL OF HUMAN REPRODUCTION

Anrudh K. Jain, University of Michigan

Introduction

To measure the fertility level in a society one can use many available measures, such as the General Fertility Rate, age specific fertility rates or duration specific birth rates. The use of a particular measure depends on the availability of data and the purpose of the study. In the absence of the required data, one often turns towards some suitable model specific to the purpose in mind. At times these models are used without adequately testing them with some observed set of data obtained from field surveys or from other sources. In this paper we shall first predict duration specific averages and variances of live births by using the Perrin and Sheps[1] stochastic model of human reproduction and then shall compare them with the corresponding observed values obtained directly from a sample of nonwestern women.

In the last few decades a number of hypothetical models for predicting fertility levels have been developed by Dandekar[2], Brass[3], Potter[4], Henry[5], and Perrin and Sheps[6]. In this paper we will confine our discussions to the stochastic model of human reproduction developed by Perrin and Sheps. This model has been applied by other investigators to study a variety of problems. For example, Sheps and Perrin[7,8] used it to study the effects of contraception on birth rates and to study the distributions of birth intervals; Sheps[9] used it to study the effect of pregnancy wastage on fertility, and Potter et. al.[10] used it to study birth interval dynamics. Sheps and Perrin have mainly used hypothetical values of the basic parameters of the model, whereas Potter et. al. used data for Indian women from the Khanna study in which the users were not separated from the nonusers of contraception. The Perrin and Sheps model, however, is properly applicable only to nonusers of contraception and we have an opportunity to compare some of the theoretical results obtained from this model with the corresponding observed values for nonusers. These comparisons do not neccessarily provide a conclusive test of the model mainly because the magnitude of divergence between the observed and theoretical values reflects not only the limitations of the model but also dificiencies in the data.

Data

The data are taken from the results of an intensive fertility survey conducted in Taichung City of Taiwan, by the Taiwan Population Studies Center, in collaboration with the Population Studies Center of the University of Michigan. The Taichung Survey[11] is based on a probability sample of 2,443 married women between the ages 20 to 39, living with their husbands, interviewed towards the end of 1962 just before a year long family planning action program.[12] Extensive information about the various demographic and socio-economic characteristics, attitudes towards family planning and practice of various family planning methods are available for each woman. In this analysis, however, only a fraction of this information is used.

The Taichung sample provides detailed information about the pregnancy histories of all women in the sample. In this sample about 84 per cent of the couples in the childbearing ages had never used contraception (excluding induced abortion) between marriage and the last pregnancy prior to interview. Even for the remaining 16 per cent who had used contraception at some time, we can use that part of their pregnancy histories in which contraception was not used. Thus, the Taichung survey provided a set of pregnancy histories free from contraception which can be used to estimate the parameters in the absence of contraception, required in the Perrin and Sheps model.

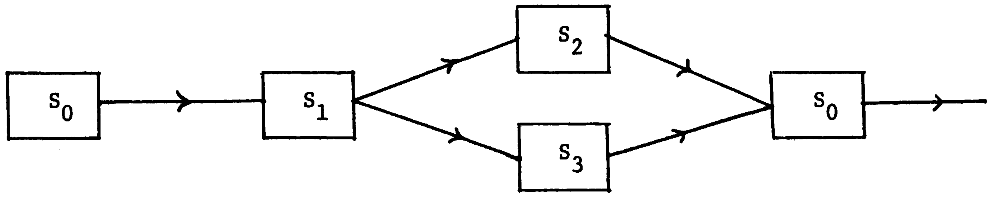
Before presenting the specific comparisons let us first review some of the basic assumptions underlying the Perrin and Sheps model and consider how well these assumptions are met by our data.

The Model[13]

The Perrin and Sheps model assumes that in a given month a woman can only be found in one of the four mutually-exclusive states: a nonpregnant, fecundable state S_0; a pregnant state S_1; an infecundable period following a pregnancy loss in state S_2; and an infecundable period following a live birth in state S_3. Here pregnancy loss includes spontaneous abortions, induced abortions, and still births, whereas in the orginal version still births were separated from spontaneous and induced abortions. The model further assumes that at the time of marriage women start in the nonpregnant fecundable state S_0, during which they are subject to some fixed probability of conception. After a random length of time they pass to state S_1. The duration of their stay in state S_1 depends upon the outcome of the pregnancy. Following the termination of a pregnancy, women pass to state S_2 or state S_3. If the pregnancy terminates in a pregnancy loss, women pass from state S_1 to state S_2; and if the pregnancy terminates in a live birth women pass to state S_3. During their stay in states S_2 or S_3 women are temporarily infecundable, i.e., the probability of conception during their stay in these states is zero. From states S_2 or S_3 women pass to state S_0 and become fecundable again, and thus the process starts once more.

In the "pure" version of the model, women are assumed to be homogeneous with respect to its parameters, namely, the moments of the

Schematically:



(The length of stay in each state, and the outcome of each pregnancy are independent random variables.)

S_0- Nonpregnant fecundable state

S_1- Pregnant State

S_2- Infecundable period following pregnancy losses

S_3- Infecundable period following live births

---

length of stay in states S_0, S_1, S_2, and S_3, and the probabilities of moving from state S_1 to state S_2 or state S_3. These parameters are assumed to be independent of women's age and the pregnancy order. The length of stay in each state and the outcome of pregnancy are assumed to be independent random variables. Moreover, the length of the reproductive period is assumed to be unlimited. In actuality as had been shown elsewhere[14], Taiwanese women in the present sample are not homogeneous, the parameters are dependent on women's age and the pregnancy order, and the reproductive period is limited. Since we are using the values of parameters estimated from real data in this "modified" version of the model, we are roughly taking into account the heterogeneity among women with respect to their age and the pregnancy order. For example, we estimated the average and variance of conception delay and of pregnancy intervals by using actual frequency distributions of all heterogeneous women observed between marriage and interview, and similarly we estimated the probability of pregnancy losses from all pregnancies during this period.[15] For applying the results of the Perrin and Sheps model, we assume that all women included in the analysis are homogeneous with respect to these estimated overall parameters, and that these overall parameters remain constant throughout the observed reproductive period.

Two assumptions, homogeneity and stability of the parameters, are only partly taken care of in this modified version of the model. The other two assumptions of the model, independence between different states and an unlimited reproductive period, are not met at all in actuality. Considering the limitations of the model, Perrin and Sheps suggested that "the period for which this model can be assumed to hold for each woman must necessarily be restricted to an interval of at most 10 to 15 years in the middle of the childbearing age."[16] Given these limitations of the model and the data, we do not expect that the observed averages and variances will agree closely with the corresponding estimated values. Nevertheless, it is useful to measure how small or great are the discrepancies between the values observed in a real situation and the values predicted by a model which has the virtue of simplicity.

Results

The Perrin and Sheps model provides approximate mathematical expressions for estimating averages and variances of the number of live births at the end of "t" months after marriage. These are shown in the appendix along with the corresponding estimation procedures.

As noted by Perrin and Sheps[17] these expressions yield good approximations to exact moments after five years of marriage, i.e., after five years of marriage the two moments of live births obtained by using approximate expressions are the same as those obtained by using exact expressions. For using these approximate expressions one requires the estimates of the first two moments of the first passage time from state S_0 to state S_3 or the first birth interval (FBI); and the first three moments of the time between the two successive visits to state S_3 or completed birth interval (BI). One can estimate these moments of FBI and BI either from the sample frequency distributions of FBI and BI or by using the results of the Perrin and Sheps model. For the latter case one needs to estimate the basic parameters of the model from the observed data. In this paper we have estimated the moments of FBI and BI by using both of these procedures. These results are compared in the appendix (Table 3).

Substituting the estimated moments of FBI and BI from the first set in expressions for the average and the variance of live births per woman after t months of marriage, we obtained two expressions: 0.0375 t - 0.1233 for average live births and 0.00602 t + 0.24210 for variance of live births. By substituting t = 1, 2, 3,...etc., we can obtain the theoretical number of live births and also their variances in one, two,... etc. months after marriage. These theoretical results will be referred as the 'Theoretical A' series.

Similarly by substituting the estimated moments of FBI and BI from the second set we obtained two expressions: 0.03590 t - 0.0888 for average live births and 0.00654 t + 0.13590 for variance of live births. The results from these expressions will be referred as the 'Theoretical B' series.

Table 1. Comparison of Observed and Theoretical Averages of Live Births per Woman in T Years of Marriage in the Absence of Contraception.

Marriage duration (years)	Observed[a/]	Theoretical[b/]		I[c/]	
		A	B	A	B
1	.302	.327	.342	8.3	13.2
2	.823	.777	.773	-5.6	-6.1
3	1.222	1.227	1.204	.4	-1.5
4	1.674	1.667	1.634	.2	-2.4
5	2.042	2.127	2.065	4.2	1.1
6	2.449	2.577	2.496	5.2	1.9
7	2.822	3.027	2.927	7.3	3.7
8	3.152	3.477	3.358	10.3	6.5
9	3.479	3.927	3.788	12.9	8.9
10	3.800	4.377	4.219	15.2	11.0
11	4.130	4.827	4.650	16.9	12.6
12	4.351	5.277	5.081	21.3	16.8
13	4.566	5.727	5.512	25.4	20.7
14	4.784	6.177	5.942	29.1	24.2
15	5.022	6.627	6.373	32.0	26.9
16	5.105	7.077	6.804	38.6	33.3
17	5.202	7.527	7.235	44.7	39.1
18	5.194	7.977	7.666	53.6	47.6
19	5.347	8.427	8.096	57.6	51.4
20	5.303	8.877	8.527	67.4	60.8

a/ Women who were premaritally pregnant or who have used contraception between marriage and last pregnancy are excluded.

b/ Theoretical average live births per woman in T years of marriage:

$$A = 0.4500\ T - 0.1233$$

$$B = 0.4308\ T - 0.0888$$

c/ $I = \frac{\text{Theoretical - Observed}}{\text{Observed}} \times 100$

We compare the above two series with the averages and variances of live births after t months of marriage obtained from the sample. If we consider women married for at least t months and their live births during these months, we frequency distributions of women by number of live births for each month after marriage. From these observed frequency distributions the averages and variances of live births can be calculated for each successive month following marriage. However, for simplicity the moments of live births are calculated only for one year intervals, i.e., for women married for at least 12, 24, 36, 48,..., etc. months. The observed and estimated values of average live births are compared in Table 1 and in Figure 1, and the corresponding values for variance of live births are compared in Table 2 and Figure 2.

The differences between the two theoretical series A and B are very small. For this reason we discuss in the following paragraphs only the differences between theoretical series B and the observed values. The differences between the observed and the predicted averages and variances of live births during the first five years of marriage could be attributed to the fact that here we used approximate expressions instead of exact expressions for predicting these averages and variances. Hence, we shall restrict our discussion to the differences after the fifth year of marriage.

As can be seen from the last column in Table 1, the differences between the observed and the predicted number of live births increase gradually with the duration of marriage. For example, for women married six years or more, the theoretical average number of live births during six years of marriage is only two percent (or about .05 births) higher than the corresponding observed average number of live births. But this difference increases gradually to 61 per cent (or about three births) for women married 20 years or more. (See Table 1 and Figure 1).

Comparisons of variances in Table 2 and Figure 2 show a different pattern. Except for the second year, the theoretical variances are always lower than the observed. For example, for women married at least six years the theoretical variance is 20 per cent (or 0.15) lower than the observed variance. This difference increases gradually to 53 per cent (or 1.9) for women married for at least 20 years. (See Table 2 and Figure 2)

These patterns in the differences could arise due to variety of reasons. Three reasons are discussed here.

Figure 1: Theoretical and Observed Number of Live Births by Duration of Marriage

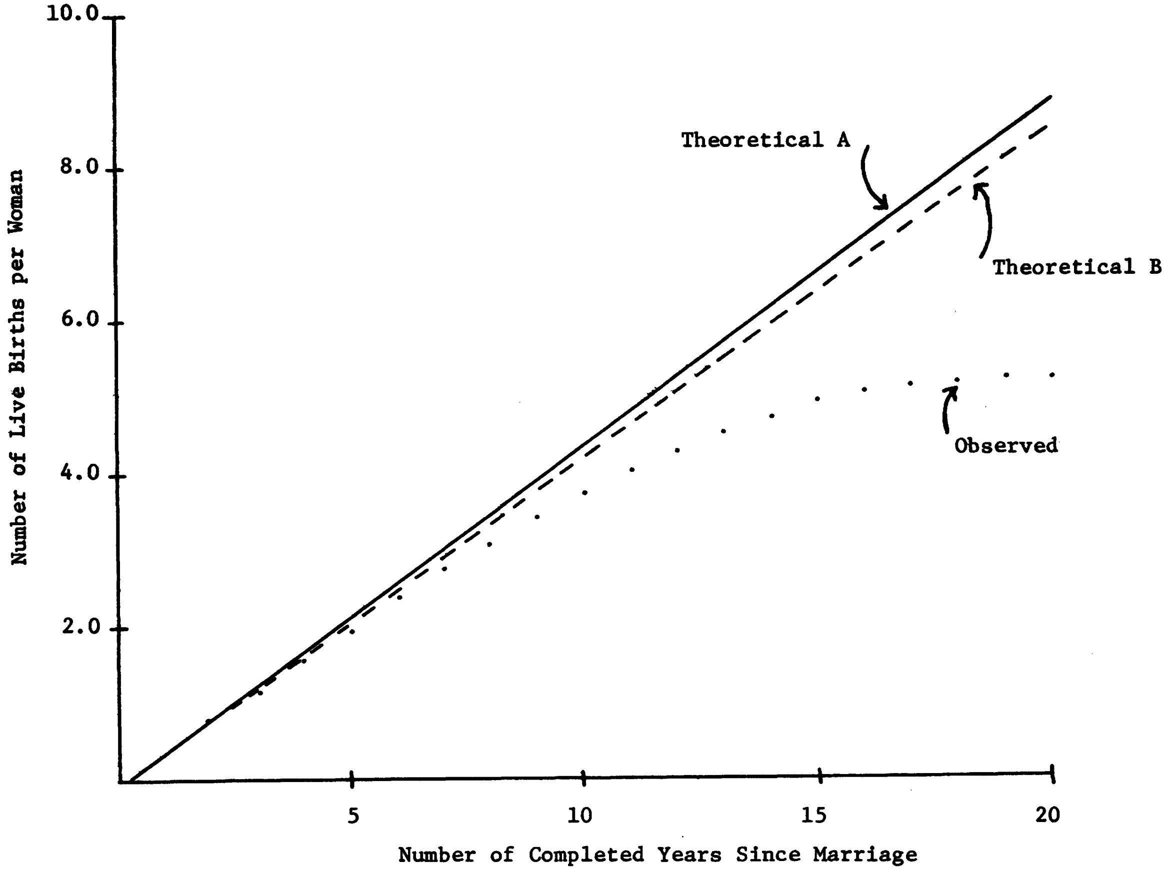


Table 2. Comparison of Observed and Theoretical Variance of Live Births per Woman in T Years of Marriage in the Absence of Contraception.

Marriage duration (years)	Observed[a/]	Theoretical[b/]		I[c/]	
		A	B	A	B
1	.215	.314	.214	46.0	-.5
2	.205	.387	.293	88.8	42.9
3	.382	.459	.371	20.2	-2.9
4	.478	.531	.450	11.1	-5.9
5	.630	.603	.528	-4.3	-16.2
6	.763	.676	.607	-11.4	-20.4
7	.928	.748	.685	-19.4	-26.2
8	1.099	.820	.764	-25.4	-30.5
9	1.343	.892	.842	-33.6	-37.3
10	1.549	.965	.921	-37.7	-40.5
11	1.766	1.037	.999	-41.3	-43.4
12	2.016	1.109	1.078	-45.0	-46.5
13	2.189	1.181	1.156	-46.0	-47.2
14	2.458	1.253	1.235	-49.0	-49.8
15	2.589	1.326	1.313	-48.8	-49.3
16	2.906	1.398	1.392	-51.9	-52.1
17	2.805	1.470	1.470	-47.6	-47.6
18	2.606	1.542	1.549	-40.8	-40.6
19	3.323	1.615	1.627	-51.4	-51.0
20	3.599	1.687	1.706	-53.1	-52.6

a/ Women who were premaritally pregnant or who have used contraception between marriage and last pregnancy are excluded.

b/ Theoretical variance of live births per woman in T years of marriage:

$$A = 0.07224\ T + 0.24210$$
$$B = 0.07848\ T + 0.13590$$

c/ $I = \frac{\text{Theoretical - Observed}}{\text{Observed}} \times 100$

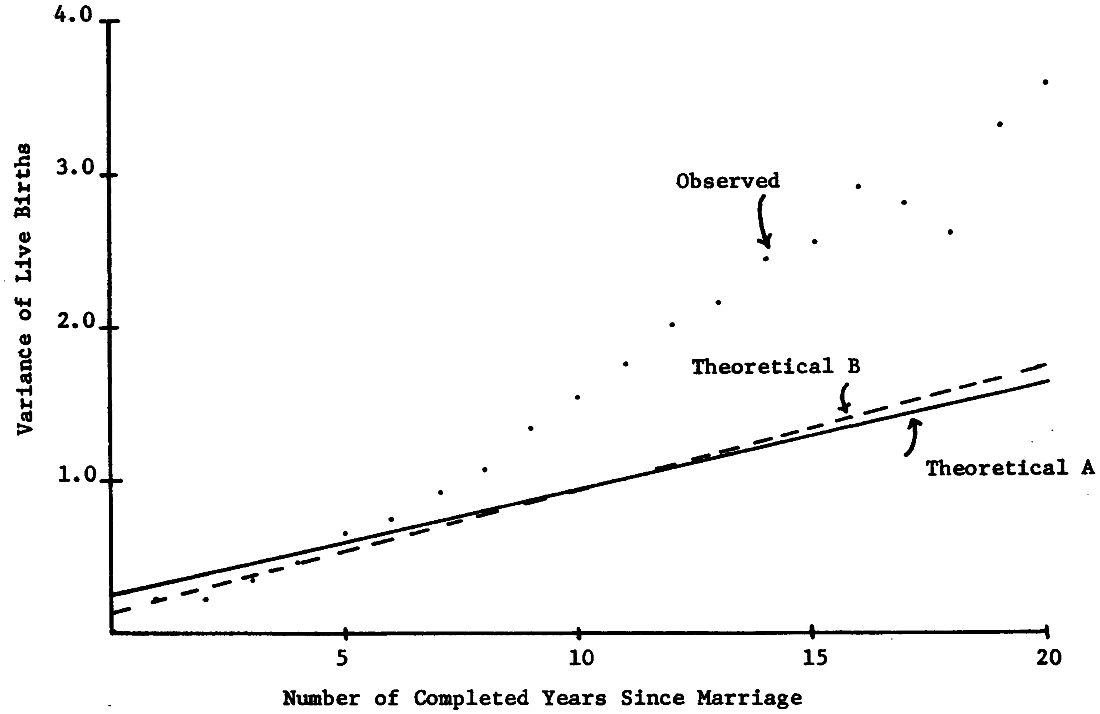
First, the apparent discrepancies between the observed and the theoretical values could be due to deficiencies in the data: for example, still births were not separated from miscarriages, and other errors were introducted due to the fact that the information was collected retrospectively. The retrospective nature of the survey might lead us to suspect some memory bias in reporting all live births by women in the sample during their childbearing period. If so, this would have deflated the observed averages and variances of live births and would have also affected the theoretical series. However, in the present sample, it is believed that the underreporting of live births or infant deaths is negligible because the births reported in the survey were compared with the births registered in the population register maintained by the Taiwanese government. Women for whom the two sources showed different numbers were reinterviewed and the discrepancies were corrected.[18]

A second reason might be the fact that the expressions relating the theoretical average and variance to the duration of marriage assume a constant rate of increase in the average and the variance of live births. Actually this is neither true for the average nor for the variance of live births. Let us consider these points separately.

Due to the increasing frequency of secondary sterility with advancing age, every year some women stop contributing more births, and some women do not produce births as quickly as they did at earlier ages. Thus, the rate of increase in the average number of live births does not remain constant: it starts declining with ascending marriage duration, and finally the rate of increase becomes zero when all women reach the end of their childbearing. In other words, the relationship between the observed number of live births and marriage duration is not linear as predicted by the Perrin and Sheps model, but rather curvilinear. This seems to be the main reason for divergence between the theoretical and observed numbers of live births.

Another consequence of the increasing prevalence of secondary sterility with advancing age is that women married for the same number of years are not homogeneous with respect to the "effective length of reproductive period," which is the period between the onset of marriage and the last live birth prior to the interview. This heterogeneity among women is mainly responsible for the divergence between the observed and the theoretical variances. If we consider women married for at least 15 years then despite their homogeneity with respect to their duration of marriage, they are not homogeneous with respect to their effective reproductive period because this group in-

Figure 2: Theoretical and Observed Variance of Live Births by Duration of Marriage



cludes subgroups of women who had their last live birth during the first, second, third,...fifteenth year after marriage. Thus, women in this group are very different with respect to their attained parity. Now as the length of marriage duration increases, the heterogeneity among women with respect to their attained parity also increases and so does the observed variance of live births. The observed variance will thus continue to increase until all women in the sample reach the end of their reproductive period estimated around 45 years. For the group of women who got married at the age of 20 years, the length of reproductive period will be about 25 years. After 25 years of marriage duration, for this group of women, the rate of increase in the observed variance of live births will be zero, whereas the theoretical variance will still continue to increase with the same constant rate of increase. The needed assumption of an unlimited length for the reproductive period, leads the observed and the theoretical curves to intersect at some point beyond usual observation range after which the theoretical variances will be higher than the observed variances for all years.

A third reason for the apparent divergence between the observed and the theoretical variances of live births might be the violation of the assumption of independence between different states of the model. For obtaining the theoretical variances, all the covariance terms are assumed to be zero; given positive association between different states one might expect positive covariances which would deflate the theoretical variances. The reasons for expecting positive covariances have been discussed elsewhere.[19] Here we will mention them briefly. Due to age dependency, a positive association can be expected between the probability of pregnancy wastage, fecundable periods, and the length of pregnancy intervals following pregnancy losses and following live births. Even with age held constant, a positive association can be expected between the lengths of fecundable and infecundable periods due to their dependences on the outcome of the preceeding pregnancy.

Out of the three reasons discussed above, the first one does not seem to be very important. However, even if one is able to collect perfect data, it seems that the values predicted by the model will not be close to the corresponding observed values because of the nonlinear relationships between marriage duration and the number of live births. The Perrin and Sheps model, in this respect, needs some modifications.

Summary

As an abstract model of fertility process the Perrin and Sheps model necessarily involves simplifying assumptions. In this paper we have compared the values predicted by the model, using parameters estimated from a real population with more variability than the model assumes, with the corresponding observed values. These comparisons show that the theoretical values of averages and variances of live births predicted by the model are not close to the corresponding observed values, even for marriage durations of 10 to 15 years in the middle of the childbearing period, during which the model was supposed to hold. These results suggest that the model's approximate expressions or the corresponding asymptotic expressions giving averages and variances of live births should be used with caution. This may be particularly true when the model is used for estimating the effects of contraception and/or induced abortion on fertility rates, and implications are drawn about the optimum distribution of contraceptive use in a population.

Acknowledgements

This paper is based on the research carried out by the author as a part of his doctoral dissertation at the University of Michigan during 1966-68 under the able guidance of Leslie Kish, Paul Demeny, Ronald Freedman, Thomas F. Mayer, and Robert G. Potter, Jr. The Taiwan Population Studies Center and the University of Michigan Population Studies Center have kindly made available to me the data from the Taichung Intensive Fertility Survey conducted under a grant from the Population Council. The author is thankful to L. P. Chow, the Director of Taiwan Population Studies Center for his cooperation. The University of Michigan Population Studies Center has provided the clerical help, and the assistance of J. Michel Coble and Claudia Ludvigh in computer programming. The author gratefully acknowledges helpful suggestions from the members of his doctoral committee and his colleagues, Beverley Duncan, Albert I. Hermalin, and James A. Palmore. Grants from the Ford Foundation and the National Institute of Child Health and Human Development provided the financial support for this research.

References

1. E. B. Perrin and M. C. Sheps, "Human Reproduction: A Stochastic Process," Biometrics, XX, No. 1 (March, 1964), 29-45.

2. V. M. Dandekar, "Certain Modified Forms of Binomial and Poission Distributions," Sankhya, XV, (1955), 211-235

3. W. Brass, "The Distribution of Births in Human Population," Population Studies, XII, No. 1 (July, 1958), 51-72.

4. R. G. Potter, "Some Relationships Between Short Range and Long Range Risks of Unwanted Pregnancy," Milbank Memorial Fund Quarterly, XXXVIII (1960), 255-263; "Length of the Fertile Period," Milbank Memorial Fund Quarterly, XXXIX (1961), 132.

5. L. Henry, "Fécundité et Famille: Modèles Mathematiques," Population, III (1957), 413-444; "Fécondité et Famille: Modèles Mathematiques II," Population, XVI (1961), 27-48 and 261-282.

6. Perrin and Sheps, op. cit.

7. M. C. Sheps and E. B. Perrin, "Changes in Birth Rates as a Function of Contraceptive Effectiveness: Some Applications of a

Stochastic Model," American Journal of Public Health, LIII (July, 1963), 1031-1045.

8. M. C. Sheps and E. B. Perrin, "The Distribution of Birth Intervals Under a Class of Stochastic Fertility Models," Population Studies, XVII, No. 3 (March, 1964), 321-331.

9. M. C. Sheps, "Pregnancy Wastage as a Factor in the Analysis of Fertility," Demography, I, No. 1 (1964), 111-118.

10. R. G. Potter, Jr., J. B. Wyon, M. Parker, and J. E. Gordon, "A Case Study of Birth Interval Dynamics," Population Studies, XIX, No. 1 (July, 1965), 81-96.

11. For details of Taichung Intensive Survey refer to A. K. Jain, Fecundity Components in Taiwan: Application of a Stochastic Model of Human Reproduction, (Unpublished Ph. D. dissertation, University of Michigan, 1968), Chapter II. Also, R. Freedman and J. Y. Takeshita, Family Planning in Taiwan: Tradition and Change, (to be published), Appendix I-3.

12. For the description of the design of the Family Planning Action Program refer to B. Berelson and R. Freedman, "A Study in Fertility Control," Scientific American, CCX, No. 5 (May, 1964), 3-11.

13. This is a brief version of the detailed description of the Perrin and Sheps model given in their original paper, op. cit.

14. Jain, op. cit.

15. For description of estimation procedure refer to Jain, op. cit.

16. Perrin and Sheps, op. cit., 33.

17. Perrin and Sheps, op. cit., 43-44.

18. For further details see Freedman and Takeshita, op. cit.

19. Jain, op. cit.

Appendix

The two expressions giving the averages and variances of live births are taken from the work of Perrin and Sheps.

The average number of live births during t months of marriage assuming that women start in the fecundable nonpregnant state S_0, is approximately:

$$E\,[LB(t)] \doteq \frac{t}{M(BI)} + \frac{S(BI)}{2[M(BI)]^2} - \frac{M(FBI)}{M(BI)} \qquad (1)$$

and the corresponding variance is approximately:

$$V[LB(t)] \doteq \frac{V(BI)}{[M(BI)]^3}\cdot t + \frac{5[S(BI)]^2}{4[M(BI)]^4} - \frac{2T(BI)}{3[M(BI)]^3} - \frac{S(BI)}{2[M(BI)]^2} - \frac{M(FBI)\,S(BI)}{[M(BI)]^3} + \frac{V(FBI)}{[M(BI)]^2} + \frac{M(FBI)}{M(BI)} \qquad (2)$$

where,

LB	Live Birth
BI	Completed Birth Interval--Period between two consecutive live births
FBI	First Birth Interval--Period between marriage and first live birth
M( )	Mean
V( )	Variance
S( )	Second moment about origin
T( )	Third moment about origin

To evaluate expressions (1) and (2) one needs to estimate the moments of the first birth interval and of completed birth intervals. The simplest way to estimate these moments is to estimate them from the oberved frequency distributions of first birth intervals and completed birth intervals. Another way to estimate these moments is to substitute the estimated values of the parameters of the model in the expressions for estimating the moments of first birth interval and completed birth interval. The procedure for estimating the parameters of the model is described in the author's Ph. D. dissertation. The estimated moments of the first birth interval and the completed birth intervals are shown in Table 3.

Table 3. Estimated Moments of First Birth Interval and Completed Birth Interval in the Absence of Contraception.

Item	Symbol	Estimated Values A	Estimated Values B
Average First Birth Interval	M(FBI)	18.76	18.91
Variance of First Birth Interval	V(FBI)	237.81	173.33
Average Birth Interval	M(BI)	26.67	27.82
Variance of Birth Interval	V(BI)	114.12	140.72
Second Moment of Birth Interval about its Origin	S(BI)	825.28	914.67
Third Moment of Birth Interval about its Orgin	T(BI)	30,991.50	36,279.78

A - Estimated from the observed frequency distributions of FBI and BI.

B - Estimated by substituting the parameters of the model in the expressions of the moments of FBI and BI.

AN EMPIRICAL STUDY OF THE STABILITIES OF ESTIMATORS AND VARIANCE ESTIMATORS IN P.P.S. SAMPLING, II

D. L. Bayless* and J. N. K. Rao

1. Introduction

In a previous paper (1967) we empirically investigated the stabilities of variance estimators jointly with the efficiencies of estimators of the population total for certain P.P.S. (Probability Proportional to Size) sampling methods for samples of n = 2 using actual finite populations and under the assumption of a super population model. In this paper after deriving the variance of the variance estimator in general and under the assumption of a super population model, we perform similar empirical studies as that mentioned above for samples of size n = 3 and 4. We also provide a computer computational scheme to calculate certain conditional probabilities for Murthy's (1957) method.

We have chosen only those methods (excepting one) which satisfy the following requirements:

a. a nonnegative, unbiased variance estimator should be available,

b. computations are feasible (timewise) on a high speed computer.

Based on these conditions we have selected the following methods for the present study:

1. The methods of Fellegi (1963), Carroll and Hartley (1964) and Sampford (1967), all using the Horvitz-Thompson (1952) estimator and satisfying $\Pi_j = np_j$ $(j = 1, \ldots, N)$, Π_j being the probability of including the j-th population unit in the sample.

2. The methods of Des Raj (1956) and Murthy (1957).

3. The method of Rao, Hartley, and Cochran (R.H.C.) (1962).

4. Lahiri's (1951) method.

The requirement a. is not satisfied by Lahiri's estimator. Nevertheless, we have included it in view of the recent work by Godambe (1966) based on concepts other than efficiency. Also, we exclude Fellegi's method for n = 4 because the computational cost becomes quite expensive due to calculation of the joint inclusion probabilities of any pair of units in the sample. Further the routine to calculate the working probabilities used to select the unit at the r-th draw, r = 1, ..., n often required several iterations for the populations we considered. Lahiri's method is excluded in our empirical study under the assumption of a super population model.

2. Formulae

In giving the formulae for the methods in this section we give the equations for any n and based on a super population model (Cochran(1946)) in which the finite population is regarded as being drawn from an infinite super population. The results obtained apply to the average of all finite populations that can be drawn from the super population. We assume the following, often used, super population model for the comparison of estimators:

$$y_i = \beta x_i + e_i \ , \ i = 1, \ldots, N$$
$$\epsilon(e_i|x_i) = 0 \ , \ \epsilon(e_i^2|x_i) = ax_i^g$$
$$\epsilon(e_ie_j|x_i,x_j) = 0 \ , \ a > 0 \ , \ g \geq 0$$

where ϵ denotes the average over all the finite populations that can be drawn from the super population. For the comparison of variance estimators we further assume that e_i's are normally distributed so that $\epsilon(e_i^4) = 3a^2x_i^{2g}$. In most practical situations, g is expected to lie between 1 and 2. Some theoretical results are available on the relative efficiencies of the estimators (Hanurav (1965), Rao (1966), and Vijayan (1967)) but no guidelines are available with regard to the relative magnitudes. Nothing is known on the stabilities of the variance estimators under the super population model.

Of course, the formulae we need for our empirical studies, and the computer programs, are those for n = 3 and 4.

The new formulae of this section are the Ev^2's while the other formulae were previously given in the references cited above. To check the formulae we considered the case when all the x-values are equal to one which is equivalent to simple random sampling. Under this condition all the formulae are identical except for Des Raj's method and were checked numerically. We also used a complete combinatorial evaluation to check the formulae for the R.H.C. method that is described in Appendix A.

2.1 Some IPPS (Inclusion Probabilities Proportional to Size) Methods Using the Horvitz-Thompson Estimator

The Horvitz-Thompson estimator of the population total, Y , for any n is

$$\hat{Y}_1 = \sum_{i=1}^{n} y_i/\Pi_i$$

where 1, 2, ..., n denote the units in the sample. For the methods of group (1) we have, since

$$\Pi_j = np_i = nx_i/\Sigma x_i \ ,$$

the Horvitz-Thompson estimators

$$\hat{Y}_1 = \sum_{i=1}^{n} y_i/np_i$$

with variance

$$V_1 = \sum\sum_{i<i'}^{N} (n^2p_ip_{i'}-\Pi_{ii'})(y_i/np_i-y_{i'}/np_{i'})^2$$

and variance estimator (due to Yates and Grundy (1953))

$$v_1 = \sum\sum_{i<i'}^{n} \left((n^2p_ip_{i'}-\Pi_{i'i'})/\Pi_{ii'}\right)(y_i/np_i-y_{i'}/np_{i'})^2$$

where $\Pi_{ii'}$ is probability of inclusion of units i and i' in the sample. Since Ev_1^2 is needed for

*Presently employed by Research Triangle Institute, Research Triangle Park, North Carolina 27709.

the variance of the variance estimator, we write

$$Ev_1^2 = \sum_s \Pi(s)v_1^2(s)$$

$$= \sum_s \Pi(s)\left[\sum\sum_{\substack{i<i' \\ s\supset i,i'}}^{n} \left((n^2p_ip_{i'}-\Pi_{ii'})/\Pi_{ii'}\right)(y_i/np_i-y_{i'}/np_i)^2\right]^2$$

where $\sum_s$ denotes summation over all $\binom{N}{n}$ possible samples, s , $v_1(s)$ is v_1 for the sample s , $\Pi(s)$ the probability of obtaining the sample s . The formulae for $\Pi_{ii'}$ and $\Pi(s)$ for the various methods can be obtained from Fellegi (1963), Carroll and Hartley (1964), and Sampford (1967) or Bayless (1968).

Substituting the super population model into V_1 above and taking expectations we have

$$\epsilon V_1 = \epsilon\left[\sum_{i<i'}^{N}\sum (n^2p_ip_{i'}-\Pi_{ii'})(e_i/np_i-e_{i'}/np_{i'})^2\right]$$

$$= \epsilon\left[\sum_i^N (e_i^2/n^2p^2p_i^2)\left(np_i(1-np_i)\right) + \sum\sum_{i\neq i'}^{N} (\Pi_{ii'}-\Pi_i\Pi_{i'})(e_ie_{i'}/n^2p_ip_{i'})\right]$$

$$= aX^g/n \sum_{i=1}^{N} (1-np_i)p_i^{g-1}$$

which is independent of $\Pi_{ii'}$. Thus, all methods that use the Horvitz-Thompson estimator with $\Pi_i = np_i$ have the save average variance.

The evaluation of ϵEv_1^2 , v_1 being the Yates-Grundy variance estimator, is obtained as follows.

Since $Ev_1^2 = \sum_s \Pi(s)\, v_1^2(s)$, where $\Pi(s)$ and $v_1(s)$ are defined above, we have

$$\epsilon Ev_1^2 = \sum_s \Pi(s)\ \epsilon v_1{'}^2(s)$$

where $v_1{'}(s)$ is $v_1(s)$ with the super population model substituted into it.

Thus, it remains to evaluate $\epsilon v_1{'}^2(s)$, we

$$\epsilon v_1{'}^2(s) = \epsilon\left[\sum_{\substack{i<i' \\ s\supset ii'}}^{n}\sum H_{ii'}(e_i/np_i-e_{i'}/np_{i'})^2\right]^2$$

where

$$H_{ii'} = (n^2p_ip_{i'}-\Pi_{ii'})/\Pi_{ii'}$$

After taking expectations, and considerable manipulation, we have

$$\epsilon v_1{'}^2(s) = \left[\sum_{\substack{i<i' \\ s\supset ii'}}\sum_{} H_{ii'}K(g)_{ii'}\right]^2 + \sum\sum_{\substack{i\neq i' \\ s\supset ii'}}^{n} H_{ii'}^2K(g)_{ii'}^2 + \sum\sum\sum_{\substack{i\neq i'\neq i'' \\ s\supset ii'i''}} H_{ii'}H_{ii''}X^{2g}/n^4p_i^4 + \sum\sum\sum_{\substack{i\neq i'\neq i'' \\ s\supset ii'i''}} H_{ii'}H_{i'i''}X_{i'}^{2g}/n^4p_{i'}^4$$

where

$$K(g)_{ii'} = X_i^g/n^2p_i^2 + X_{i'}^g/n^2p_{i'}^2 \ .$$

2.2 The Des Raj Method

Des Raj proposed the uncorrelated unbiased estimators

$$t_1 = y_1'/p_1'$$

$$\cdots$$

$$t_r = \sum_{t-1}^{r-1} y_t' + y_r'(1-\sum_{t=1}^{r-1} p_t')/p_r'$$

$$\cdots$$

$$t_n = \sum_{t=1}^{n-1} y_t' + y_n'(1-\sum_{t=1}^{n-1} p_t')/p_n'$$

where (y_r',p_r') denote the y-value and the p-value of the unit selected at the r-th draw. As an unbiased estimator of Y we have

$$\hat{Y}_2 = \bar{t} = (1/n)\sum_{i=1}^{n} t_i$$

with

$$V_2 = V(\hat{Y}_2) = \frac{1}{2n^2}\sum_{i<i'}^{N}\sum p_ip_{i'}\left[1+\sum_{r=2}^{n} Q_{ii'}(r-1)\right](y_i/p_i - y_{i'}/p_{i'})^2$$

where $Q_{ii'}(r-1)$ denotes the probability of non-inclusion of unit i and i' in the first $r-1$ sample units and $Q_{ii}(0) = 1$ and variance estimator

$$v_2 = \Sigma(t_i-\bar{t})^2/n(n-1) \ .$$

Since the Q_{ii}'s of V_2 are very cumbersome to calculate, the formulae we used to calculate V_2 and Ev_2^2 using the above notation and letting $v_2(s')$ be v_2 for one of the $n!$ possible orderings, s' , of the sample s , are

$$V_2 = \sum_{ss'}\sum p_2(s')v_2(s')$$

and

$$Ev_2^2 = \sum_{ss'}\sum p_2(s')v_2^2(s') .$$

where $\sum_s$ and $\sum_s$ denote the summation over all possible $\binom{N}{n}$ samples of size n and all possible n! orderings of a given sample, s , of size n respectively and

$$p_2(s') = p_{\ell_1}\left[p_{\ell_2}/(1-p_{\ell_1})\ldots p_{\ell_n}/(1-p_{\ell_1}-\ldots p_{\ell_{n-1}})\right]$$

$$(s'\supset \ell_t \quad t = 1, \ldots, n) .$$

Substituting the super population into $v_2(s)$ to obtain $v_2'(s)$, we have $\epsilon V_2'$ and $\epsilon Ev_2'^2$ as defined as

$$\epsilon V_2' = \sum_{ss'}\sum p_2(s') \, \epsilon v_2'(s')$$

and

$$\epsilon Ev_{22}'^2 = \sum_{ss'}\sum p_2(s') \, \epsilon v_2'^2(s')$$

In appendix E we derive $\epsilon v_2'(s')$ and $\epsilon v_2'^2(s')$.

2.3 The Murthy Method

Murthy's estimator of Y for any n is

$$\hat{Y}_3 = \sum_{i=1}^{n} p(s|i)y_i/p(s)$$

where p(s) denotes the probability of getting the sample of n units; p(s|i) denotes the conditional probability of getting s given that unit "i" was drawn first (i = 1, 2, ..., n) .

Hence, for Murthy's selection procedure we have

$$p(s) = p_3(s) = \sum_{s'} p_3(s')$$

where $p_3(s')$ is the same as $p_2(s')$ of section 2.2 and $\sum_{s'}$ is defined in section 2.2. A scheme to calculate the p(s|i)'s is given in Appendix D.

The variance of Murthy's estimator is

$$V_3 = V(\hat{Y}_3) = (1/2)\sum_{ii'}^{N}\sum\left\{(y_i/p_i-y_{i'}/p_{i'})^2 \right.$$

$$\left. p_ip_{i'}\left[1-\sum_{s\supset ii'}^{*} p(s|i)p(s|i')/p(s)\right]\right\}$$

where $\sum^{*}_{s\supset ii'}$ denotes summation over all sample s that contain units i and i' , with variance estimator

$$v_3 = v(\hat{Y}_3) = 1/2\sum_{i\neq i}^{n}\sum\left\{p_ip_{i'}[p(s)p(s|ii') - p(s|i) \right.$$

$$\left. p(s|i')]/p^2(s)(y_i/p_i-y_{i'}/p_{i'})^2\right\}$$

where p(s|ii') denotes the conditional probability of s given that i and i' have been selected in the first 2 draws. In Appendix D we give a computing scheme to calculate the p(s|ii')'s .

Since V_3 involves the cumbersome sum $\sum^{*}_{s\supset ii'}$ we use a different formula to calculate V_3 which is easier to compute on a computer. It is

$$V_3 = \sum_s p_3(s)v_3(s)$$

where $v_3(s)$ is v_3 for the sample s . Also,

$$Ev_3^2 = \sum_s p_3(s)v_3^2(s) .$$

Substituting $y_i = \beta x_i + e_i$ into $v_3(s)$ to obtain $v_3'(s)$ we have ϵV_3 and ϵEv_3^2

$$\epsilon V_3 = \sum_s p_3(s) \, \epsilon v_3'(s)$$

$$\epsilon Ev_3^2 = \sum_s p_3(s) \, \epsilon v_3'^2(s) ,$$

where

$$\epsilon v_3(s) = \sum_{\substack{i<i' \\ s\supset ii'}}^{n}\sum M_{ii'} \, K(g)_{ii'}$$

with $K(g)_{ii'}$ defined in section 2.2 and

$$M_{ii'} = p_ip_{i'}[p(s)p(s|ii')-p(s|i)p(s|i')]/p^2(s)$$

and

$$\epsilon v_3'^2(s) = [\epsilon v_3'(s)]^2 + \sum_{\substack{i\neq i' \\ s\supset ii'}}^{n}\sum M_{ii'}^2 K(g)_{ii'}^2$$

$$+ \sum_{\substack{i\neq i'\neq i'' \\ s\supset ii'i''}}\sum\sum M_{ii'}M_{ii''} \, x_i^{2g}/n^4p_i^4$$

$$+ \sum_{\substack{i\neq i'\neq i'' \\ s\supset ii'i''}}\sum\sum M_{ii'}M_{i'i''} \, x_{i'}^{2g}/n^4p_{i'}^4$$

2.4 The R.H.C. Method

The R.H.C. estimator of Y for any n is

$$\hat{Y}_4 = \sum_{i=1}^{n} y_iG_i/p_i$$

where $G_i = \sum_{Gr.i} p_t$ (i = 1, ..., n) and $\sum_{Gr.i}$ denotes summation over the p-values in random group i. The variance and variance estimator of $\hat{Y}_4$ are given by

$$V_4 = K\left(\sum_{t=1}^{N} y_t^2/np_t-Y^2/n\right)$$

where

$$K = n(\Sigma N_i^2-N)/N(N-1)$$

and

$$V_4 = W\left[\Sigma G_i (y_i/p_i - \hat{Y}_4)^2\right]$$

where

$$W = (\Sigma N_i^2 - N)/(N^2 - \Sigma N_i^2) \ .$$

In Appendix A a derivation of Ev_4^2 is given for any n. By substituting $y_i = \beta x_i + e_i$ into V_4 and taking the expectation, we have

$$\varepsilon V_4 = K\left[\Sigma x_t^g/p_t - \Sigma x_t^g\right]$$

where

$$K = (\Sigma N_i^2 - N)/N(N-1)a \ .$$

A derivation of $\epsilon E v_4^2$ is given in Appendix C.

2.5 The Lahiri Method

Lahiri's estimator of Y for any n is

$$\hat{Y}_5 = \sum_{i=1}^{n} y_i / \sum_{i=1}^{n} p_i \ .$$

Using Lahiri's selection procedure it is easy to show that

$$p_5(s) = \sum_{s \supset i} x_i / \binom{N-1}{n-1} X$$

is the probability of obtaining the sample s. Hence, the variance of $\hat{Y}_5$ is

$$V_5 = \sum_s p_5(s)\, \hat{Y}_5^2(s) - Y^2$$

where $\hat{Y}_5(s)$ is $\hat{Y}_5$ for the sample s with variance estimator

$$v_5 = v_5(s) = \hat{Y}_5^2(s) - \left[\left(\sum_{s \supset i}^{n} y_i^2\right) / \binom{N-1}{n-1} p_5(s) + \left(\sum_{\substack{s \supset i i' \\ i \neq i'}}^{n}\sum y_i y_{i'}\right) / \binom{N-2}{n-2} p_5(s)\right]$$

where we note the expression in brackets is an unbiased estimate of Y^2 . Also

$$Ev_5^2 = \sum_s p_5(s)\, v_5^2(s) \ .$$

By the modified Lahiri variance estimator we mean

$$v_5' = \begin{cases} 0 & \text{when } v_5 \leq 0 \\ v_5 & \text{when } v_5 > 0 \end{cases} \ .$$

Thus, we have the mean square error as

$$\begin{aligned} \text{M.S.E.}(v'_5) &= E(v_5' - V_5)^2 \\ &= Ev_5'^2 - 2V_5 Ev_5' + V_5^2 \end{aligned}$$

where $Ev_5'^2$ and Ev_5' are easily obtained using $p_5(s)$ and the definition of v_5' .

2.6 The With Replacement Method

The customary estimator in unequal probability sampling and with replacement for any n is

$$\hat{Y}_6 = \frac{1}{n}\Sigma\, y_i/p_i$$

$$V_6 = \sum_{t=1}^{N} y_t^2/np_t - Y^2/n$$

and variance estimator

$$v_6 = \sum_{t=1}^{n} (y_t/p_t - \hat{Y}_6)^2/n(n-1) \ .$$

A derivation of Ev_6^2 for any n is given in Appendix B.

3. Empirical Results for n = 3

3.1 The Populations

We have chosen 14 natural populations for the empirical study described in Table 3.1. The first 12 populations were in the study for n = 2 (Rao and Bayless (1967)) with some of the populations sizes, N , reduced so as to reduce the amount of calculation. For example, we have included only one of the four Hanurav (1967) populations we considered in the n = 2 empirical study because they gave practically the same efficiencies for all methods. Similar reasoning was used for the omission of the other populations. We have added two natural populations from Yates' (1960) textbook, natural populations 13 and 14.

In observing Table 3.1 we see that the population sizes, N , range from 10 to 20, coefficient of variation of x, C.V.(x), range from .14 and 1.06, and correlations, ρ , from .50 to .99. Natural population 9 is different from the others in that it contains one large y/x ratio.

3.2 Stabilities of the Estimators

Table 3.2 gives the percent gains in efficiency of the estimators over Sampford's estimator for n = 3 (i.e., (V(Sampford's est.)/V(est.)-1) x 100) for the populations of Table 3.1. The following conclusions may be drawn:

1. The efficiencies of the Carroll-Hartley and Sampford estimators are about the same while Fellegi's estimator is consistently less efficient than either of them.
2. Des Raj's estimator is more efficient than the R.H.C. estimator, except for the small losses by populations 4, 5, 7, 9, and 12.
3. The loss in efficiency of Des Raj's estimator over Murthy's estimator is consistently small, with no differences in efficiencies for any given population being greater than five percentage points.
4. As with samples of size two, Lahiri's estimator is considerably more efficient than the other estimators when one or two units in the population have large sizes relative to the sizes of the remaining units, and samples containing these units have y-values that give good estimators of the population total Y(viz. populations 8 and 9). Otherwise, Lahiri's estimator has very poor efficiency compared with the other estimators. It looses to the customary estimator in p.p.s. sampling and with replacement in 5 of the 14 populations.
5. Murthy's estimator is consistently more efficient than the R.H.C. estimator except for small loss of one percentage point for population 9.
6. Murthy's estimator is slightly better than those of Sampford, Carroll-Hartley, and

Fellegi, the only loss of any magnitude being for population one.

3.3 Stabilities of the Variance Estimators

The measure of stability used to compare the variance estimators under study is (C.V.^{2}(Sampford's Var. Est.)/C.V.^{2}(Var. Est.)-1) x 100 . Table 3.3 gives the percent gains in efficiency of the variance estimators over Sampford's variance estimator for the populations listed in Table 3.1. The conclusions we draw are as follows:

1. Lahiri's modified variance estimator is consistently, and considerably, less efficient than the other variance estimators in the study except for population 9 with the 'wild' y/x ratio where it has a striking 71 percent gain over Sampford's estimator.
2. Stabilities of the Sampford, Carroll-Hartley, and Fellegi variance estimators are essentially the same for all populations.
3. The variance estimators of Murthy, Des Raj, and R.H.C. are consistently better than the "with replacement" estimator.
4. The R.H.C. variance estimator is more efficient than the Sampford, Carroll-Hartley, Fellegi, and Des Raj variance estimators for all populations excepting populations 5 and 6.
5. Murthy's variance estimator is consistently more efficient than Des Raj's variance estimator; however, the gains are small. Murthy's and Des Raj's variance estimators are more efficient than those of Sampford, Carroll-Hartley, and Fellegi excepting for population 1.
6. The R.H.C. variance estimator is often more efficient than Murthy's variance estimator.

3.4 Stabilities of the Estimators under the Assumption of the Super Population Model

Table 3.4 gives the percent gains in average efficiency of the estimators over Sampford's estimator (i.e., (ϵV(Sampford's est.)/ϵV(est.)-1) x 100) for the populations of Table 3.1 for g = 1.50, 1.75, and 2.00. Since g is usually expected to be ≥ 1.5, we have not included values of $g < 1.5$ to save computer time. The conclusions we draw are as follows:

1. As in the case n = 2, Murthy's estimator is always more efficient than the Horvitz-Thompson estimator for $g \leq 1.75$ over all populations. For g = 2 Murthy's estimator looses by no more than two percentage points to the Horvitz-Thompson estimator.
2. For all values of g considered, Murthy's estimator is consistently more efficient than the R.H.C. estimator. For some populations the gains are considerable.
3. Des Raj's estimator is less efficient than the Sampford estimator for $g \geq 1.5$.
4. Des Raj's estimator is more efficient than the R.H.C. estimator for all populations and all values of g except for populations 5 and 7 where a loss of less than one percentage point results.

3.5 Stabilities of the Variance Estimators under Assumption of the Super Population Model

The most appropriate measure of the stability of a variance estimator v under the assumption of the super population model appears to be $\epsilon[\text{C.V.}^2(v)]$, i.e., average (C.V.)^2 of the variance estimator. However, since $\epsilon[\text{C.V.}^2(v)]$ is the expectation of the ratio of two random variables, the evaluation is difficult. We have, therefore, used the alternative measures

$$\frac{\epsilon E[v-\epsilon V]^2}{(\epsilon V^2)} = \frac{\epsilon[Ev^2] - (\epsilon V)^2}{(\epsilon V)^2}$$

which is readily evaluated. Notice that this measure actually measures the variability of v around the average variance ϵV . We could also have considered the measure $\epsilon V(v)/(\epsilon V)^2$. We, however, expect that the measures $\epsilon V(v)/(\epsilon V)^2$ and $\epsilon[\text{C.V.}^2(v_i)]$ would lead to similar conclusions. It will be seen that the above measures are independent of β for all the methods considered here.

Using the stability measure above, we present in Table 3.5 the percent gains in average efficiency of the variance estimators over Sampford's estimator for g-values of 1.50, 1.75, and 2.00 for the populations of Table 3.1. The conclusions we draw are as follows:

1. The R.H.C. variance estimator is consistently more efficient than Murthy's variance estimator for $g \leq 1.75$, except for g = 1.50 and population 7. However, for g = 2, Murthy's variance estimator is consistently more efficient than the R.H.C. variance estimator.
2. The absolute percent differences between Murthy's variance estimator stability and Des Raj's variance estimator stability is less than or equal to 3 percent for all populations except population 4 for $g \leq 1.75$, and populations 4 and 8 for g = 2.00. Thus, these variance estimators are of about the same stability.
3. The R.H.C. and Murthy's variance estimators are consistently more efficient than Sampford's variance estimator as well as Carroll-Hartley, and Fellegi's variance estimator for all g. The Des Raj variance estimator is also consistently more efficient except for a few very small losses. The gains are appreciable for several of the populations.
4. The stabilities of Carroll-Hartley, and Sampford variance estimators are practically identical for all values of g.
5. Fellegi's variance estimator is consistently less efficient compared to the Sampford variance estimator for all values of g. For some of the populations, the losses are large.

4. Empirical Results for n = 4

4.1 The Populations

For this empirical study, we have selected 10 populations out of the 14 populations used for n = 3, with some of the populations sizes, N , decreased for computational reasons with the units making up the populations selected at random.

Table 4.1 describes the 10 populations where we see that N ranges from 10 to 16, C.V.(X) from .14 to 1.06, and ρ from .65 to .99. Population 7 corresponds to population 9 of Table 3.1 for n = 3 in that it contains the one large y/x ratio.

4.2 Stabilities of the Estimators

The percent gains in efficiency of the estimators over Sampford's estimator for n = 4, for the populations listed in Table 4.1, are given in Table 4.2. The conclusions that we draw are as follows:

1. The efficiencies of the Carroll-Hartley and the Sampford estimators are practically the same.
2. Murthy's estimator is consistently more efficient than the R.H.C. estimator.
3. The losses of the Des Raj estimator over Murthy's estimator are large for certain populations; in particular, a difference of at least 7 percentage points exists for populations 3, 4, 6, 7, and 9. A reason for this is that the n/N is large, at least 25%, and/or C.V.(X) is large. Although Pathak (1967) proved that if N is large compared to n , the variance of the Des Raj and Murthy estimators are identical to O(N'). It appears N is not large enough for his theory to apply in this case.
4. The Des Raj estimator appears slightly more efficient than the R.H.C. estimator.
5. In 4 out of the 10 populations, Lahiri's estimator has better efficiency than the other estimators, viz. populations 2, 4, 6, 7. The apparent reason for this is due to the small C.V.(X)-values for populations 2, 4, and 6 and the large variations in the y/x values for population 7. Since we know that, as C.V.(X)$\to$0 , all these methods tend to equal probability sampling without replacement; we would expect Lahiri's estimator to have good efficiency for small C.V.(X)-values. For most of the other populations, Lahiri's estimator has very poor efficiency compared to the other estimators. The customary estimator in p.p.s. sampling and with replacement has better efficiency than Lahiri's estimator for populations 3 and 5.

As with n = 2 and 3, it still appears that Murthy's estimator compares favorably with the Carroll-Hartley and Sampford estimators.

4.3 Stabilities of the Variance Estimators

The measure of stability used to compare the variance estimators is the same as that used for n = 3 in section 3.3. Table 4.3 gives the percent gains in efficiency of the variance estimators over Sampford's variance estimator for the populations of Table 4.1. The conclusions that we draw are as follows:

1. The Carroll-Hartley variance estimator and Sampford's variance estimator are essentially the same with regard to stability.
2. The variance estimators of Murthy, Des Raj, and R.H.C. are consistently more efficient than Sampford's variance estimator or the customary variance estimator in sampling with replacement.
3. Murthy's variance estimator is more efficient than the Des Raj variance estimator excepting that for populations 8 and 9 the latter is slightly better.
4. The R.H.C. variance estimator is more efficient than Murthy's variance estimator except for the small losses for populations 1 and 5.
5. The Modified Lahiri variance estimator has essentially the same efficiency as the unbiased Lahiri variance estimator except for population 7 which has the one large y/x ratio. These variance estimators are considerably less efficient than are the other variance estimators.

4.4 The Populations under the Assumption of the Super Population Model

Table 4.4 gives the populations for the empirical study under consideration. For all populations we have chosen the population sizes, N , to be 10 with C.V.(x) values ranging from .14 to .82 and ρ from .34 to .99. The populations were selected at random from the original populations subject to the restriction that the condition $\Pi_i \leq 4p_i$ is satisfied. Of course, the reason for choosing N = 10 for all the populations is due to computational costs. It should be pointed out that, due to the sampling fraction of 40%, comparisons of efficiency from n = 3 to n = 4 are not meaningful.

4.5 Stabilities of the Estimators under the Assumption of the Super Population Model

The Table 4.5 gives the percent gains in average efficiency of the estimators over Sampford's estimator for the populations of Table 4.4 for g = 1.5, 1.75, and 2.00. The following conclusions are drawn:

1. Except for the one percent loss in efficiency for populations 4 and 6 with small C.V.(y) and C.V.(x), Des Raj's estimator is slightly more efficient than the R.H.C. estimator for all g-values and populations.
2. As with n = 2 and 3, Murthy's estimator is more efficient than Sampford's estimator for all populations when g < 1.75. For g = 2, Murthy's estimator looses 8 and 9 percentage points in efficiency to Sampford's estimator for the populations 3 and 5 with highest C.V.(y) and C.V.(x).
3. The gains in efficiency of Murthy's estimator over Des Raj's estimator are considerable, expecially for the populations with large C.V.(x). A reason for this is that the sampling fraction is large for this study.

4.6 Stabilities of the Variance Estimators under the Assumption of the Super Population Model

Using the stability measures defined in section 3.5, Table 4.6 gives the percent gains in average efficiency for g = 1.50, 1.75, and 2.00 for the populations of Table 4.4. The following conclusions are drawn:

1. For all g-values and populations the stability of Carroll-Hartley variance estimator and the Sampford variance estimators are practically the same except for population 3 with a large C.V.(x) value.
2. For g = 1.5 Murthy's variance estimator is less efficient than the R.H.C. variance

estimator, except for the small gains for populations 4 and 6. For $g = 1.75$ it is not clear cut because Murthy's variance estimator is more efficient for 4 of the 10 populations with the losses much smaller than that in the case of $n = 3$. For $g = 2.0$, Murthy's variance estimator is consistently more efficient than the R.H.C. estimator for all populations. This conclusion agrees with our results for $n = 3$ but not with those for $n = 2$.

3. Murthy's variance estimator is consistently more efficient than Des Raj's variance estimator for all g- values and all populations.
4. The variance estimators of Murthy, Des Raj, and R.H.C. are consistently more efficient than those of Carroll-Hartley and Sampford for all populations and g-values except for the very small losses of the Des Raj's variance estimator for populations 6 (with the $g = 2$). The gains in efficiency are considerable for several of the populations, especially for those with moderate or large C.V.(x).

5. Overall Conclusions

The following overall ($n = 3$, and 4) conclusions may be drawn from our studies:

1. It appears the results under the super population model are in agreement with those from the empirical study using the actual y-data.
2. The I.P.P.S. sampling methods using the Horvitz-Thompson estimator are practically the same with respect to efficiencies of estimating the population total and the stabilities of variance estimation.
3. Murthy's method is preferable over the other methods when a stable estimator as well as a stable variance estimator are required.
4. The R.H.C. variance estimator is fairly stable, but the R.H.C. estimator might lead to significant losses in efficiency.

LIST OF REFERENCES

Bayless, D. L. 1968. "Variance estimation in sampling from finite populations," Unpublished Ph.D. thesis, Texas A&M University, College Station, Texas.

Carroll, J. L., and Hartley, H. O. 1964. "The symmetric method of unequal probability sampling without replacement," Unpublished to date, abstract in Biometrics (1964), 20, pp. 908-9.

Cochran, W. G. 1946. "Relative accuracies of systematic and stratified random samples for a certain class of populations," Annals Mathematical Statistics, 17, pp. 164-177.

______. 1963. Sampling Techniques, 2nd Edition, New York: John Wiley and Sons.

Des Raj. 1956a. "Some estimators in sampling with varying probabilities without replacement," Journal of the American Statistical Association, 51, pp. 269-84.

______. 1965. "Variance estimation in randomized systematic sampling with probability proportional to size," Journal of the American Statistical Association, 60, pp. 278-84.

Fellegi, I. P. 1963. "Sampling with varying probabilities without replacement: rotating and nonrotating samples," Journal of the American Statistical Association, 58, pp. 183-201.

Godambe, V. P. 1966. "A new approach to sampling from a finite population-I," Journal of the Royal Statistical Society, Series B, 28, pp. 310-19.

Hanurav. T. V. 1965. "Optimum sampling strategies and some related problems," Unpublished Ph.D. dissertation. Library, Indian Statistical Institute, Calcutta, India.

______. 1967. "Optimum utilization of auxiliary information: ΠPS sampling of two units from a stratum," Journal of the Royal Statistical Society, Series B, 29, pp. 374-91.

Horvitz, D. G., and Thompson, D. J. 1952. "A generalization of sampling without replacement from a finite universe," Journal of the American Statistical Association, 47, pp. 663-85.

Kish, L. 1965. Survey Sampling, New York: John Wiley and Sons.

Lahiri, D. B. 1951. "A method for sample selection providing unbiased ratio estimates," International Statistical Institute Bulletin, 33, No. 2, pp. 133-140.

Rao, J.N.K. and Bayless, D. L. "An empirical study of the stabilities of estimators and variance estimators in PPS sampling," paper presented at American Statistical Association Meeting, Proceedings of the Social Statistics Section held in Washington, D. C., December 27-30, 1967.

Sampford, M. R. 1962. An introduction to sampling theory, Edinburgh and London: Oliver and Boyd Ltd.

______. 1967. "On sampling without replacement with unequal probabilities of selection," Biometrica, 5, pp. 499-513.

Sukhatme, P. V. 1954. Sampling theory of surveys with applications, Ames, Iowa: Iowa State College Press.

Vijayan, K. 1966. "On Horvitz-Thompson and Des Raj estimators," Sankhyā, Series A, 28, pp. 87-92.

Yates, F. 1960. Sampling methods for census and surveys, Charles Griffin and Company, Inc., London.

Yates F., and Grundy, P. M. 1953. "Selection without replacement from within strata with probability proportional to size," Journal of the Royal Statistical Society, Series B, 15, pp. 253-61.

APPENDIX A

A derivation of Ev_4^2 for the R.H.C. method for any n

The R.H.C. sampling scheme involves splitting the population at random into n groups of sizes $N_1, \ldots, N_n$ where $N_1 + \ldots + N_n = N$. Then draw a sample of size one with probabilities proportional to $p_t = x_i/\Sigma x_t$ from each of these n groups independently. Thus, if the t-th unit falls in group i, the probability that it will be selected is p_t/G_i where $G_i = \sum_{\text{Group } i} p_t$. With this 'set up'

$\hat{Y}_4 = \sum_{i=1}^{n} y_i \frac{G_i}{p_i}$ is an unbiased estimate of Y with variance $V_4 = k_1 V_6$ where $k_1 = n(\Sigma N_i^2 - N)/N(N-1)$ and V_6 is the well-known formula for the variance of the customary estimator in sampling with unequal probabilities and with replacement. Also

$$v_4 = k_2 \Sigma G_i\left(\frac{y_i}{p_i} - \hat{Y}_4\right)^2$$

, can be shown to be an unbiased estimate of V_4, where

$$k_2 = (\Sigma N_i^2 - N)/(N^2 - \Sigma N_i^2) .$$

Thus, in order to derive $V(v_4)$ we need to determine Ev_4^2. This will be done by using the familiar conditional expectation argument. Let E_2 denote the expectation for a given split of the population and E_1 the expectation over all possible splits of the population into n groups of sizes $N_1, \ldots, N_n$. Therefore $Ev_4^2 = E_1E_2v_4^2$.

Now

$$v_4^2 = k_2^2\left[(\Sigma G_iy_i^2/p_i^2)^2 - 2(\Sigma G_iy_i^2/p_i^2)(\Sigma G_iy_i/p_i)^2 + (\Sigma y_iG_i/p_i)^4\right]$$

and taking expectation with respect to E_2 term by term of v_4^2 and introducing the indicator random variable

$$t_{i\ell} = \begin{cases} 1 \text{ if the } \ell\text{-th unit falls in the i-th group} \\ 0 \text{ otherwise} \\ i = 1, \ldots, n \text{ and } \ell = 1, \ldots, N \end{cases}$$

where

$$\sum_{i=1}^{n} t_{i\ell} = 1 \qquad \sum_{\ell=1}^{N} t_{i\ell} = N_i$$

and letting

$$C_{inm} = \sum_{\ell=1}^{N} t_{i\ell} y_\ell^n / p_\ell^m$$

we have

$$D_1(\underline{Y},\underline{p},\underline{T}) = E_2(\Sigma G_iy_i^2/p_i^2)^2 = \sum_{i=1} C_{i,4,3}\, C_{i,0-1}$$

$$+ \sum\sum_{i\neq i'}^{n} C_{i,2,1}\, C_{i',21}$$

$$D_2(Y,p,T) = E_2(\Sigma G_iy_i^2/p_i^2)(\Sigma G_iy_i/p_i)^2 = \sum_{i=1}^{n} C_{i,4,3}$$

$$C_{i,0,-1}^2 + \sum\sum_{i\neq i'}^{n} C_{i,2,1}\, C_{i',2,1}\, C_{i',0,-1}$$

$$+ 2 \sum\sum_{i\neq i'}^{n} C_{i,0,-1}C_{i,3,2}C_{i',1,0}$$

$$+ \sum\sum\sum_{i\neq i'\neq i''}^{n} C_{i,2,1}C_{i',1,0}C_{i'',1,0}$$

$$D_3(\underline{Y},\underline{p},T) = E_2\left(\sum_{i=1}^{n} y_iG_i/p_i\right)^4 = \sum_{i=1}^{n} C_{i,0,-1}^3 C_{i,4,3}$$

$$+ 3 \sum\sum_{i\neq i'}^{n} C_{i,0,-1}C_{i,2,1}C_{i',2,1}C_{i',0,-1}$$

$$+ 4 \sum\sum_{i\neq i'}^{n} C_{i,0,-1}^2C_{i,3,2}C_{i',1,0}$$

$$+ 6 \sum\sum\sum_{i\neq i'\neq i''}^{n} C_{i,0,-1}C_{i,2,1}$$

$$\cdot\, C_{i,1,0}C_{i'',1,0}$$

$$+ \sum\sum\sum\sum_{i\neq i'\neq i''\neq i'''}^{n} C_{i,1,0}C_{i',1,0}$$

$$\cdot\, C'_{i'',1,0}C_{i''',1,0}$$

where $\underline{Y} = \left[y_1, \ldots, y_N\right]$, is N x 1 vector of y-value;

$\underline{p} = \left[p_1, \ldots, p_N\right]$, is N x 1 vector of p-value;

$T = \{t_{i\ell}\}$ is n x N matrix of the indicator random variables $t_{i\ell}$. Hence,

$$Ev_4^2 = E_1D_1(\underline{Y},\underline{p},T) - 2E_1D_2(\underline{Y},\underline{p},T) + E_1D_3(\underline{Y},\underline{p},T).$$

Taking expectation using E_1 is quite involved and before doing so we might point that an easy combitorial solution to evaluating Ev_4^2 would be to enumerate all the possible

$$W = \frac{N!}{N_1!\ldots N_n!} = \binom{N}{N_1}\binom{N-N_1}{N_2}\cdots\binom{N-N_1\ldots N_n}{N_n}$$

distinct T matrices, each one corresponding to a different 'split' of the population, and for each

T evaluate $D_1(\underline{Y},\underline{p},T)$, $D_2(\underline{Y},\underline{p},T)$, and $D_3(\underline{Y},\underline{p},T)$ and then take the simple average of all W evaluations. Of course, this solution is much simpler than taking E_1 but the computation becomes out of practical reach when either n or N is large. For small values of n and N this approach was used to check on the solution below.

To evaluate $E_1D_j(\underline{Y},\underline{p},T)$ j = 1, 2, 3 the following notation and conditional expectation or probabilities are needed.

Letting $A_{ij} = \sum_{\ell=1}^{N} y_\ell^i/p_\ell^j$ and $b_{ijk} = \frac{N_i-j}{N-k}$

and from equal probability sampling and without replacement we have

$$E_1(t_{i\ell}) = p_r(t_{i\ell}=1) = N_i/N = b_{i,0,0}$$

$$E_1(t_{i\ell}t_{i\ell'}) = b_{i,0,0}b_{i,1,1}$$

$$E_1(t_{i\ell}t_{i\ell'}t_{i\ell''}) = b_{i,0,0}b_{i,1,1}b_{i,2,2}$$

$$E_1(t_{i\ell}t_{i\ell'}t_{i\ell''}t_{i\ell'''}) = b_{i00}b_{i11}b_{i22}b_{i33}$$

$$E_1(t_{i\ell}t_{i'\ell'}) = b_{i00}b_{i'01}$$

$$E_1(t_{i\ell}t_{i\ell'}t_{i'\ell''}) = b_{i,0,0}b_{i,1,1}b_{i',0,2}$$

$$E_1(t_{i\ell}t_{i'\ell'}t_{i''\ell''}) = b_{i,0,0}b_{i',0,1}b_{i'',0,2}$$

$$E_1(t_{i\ell}t_{i'\ell'}t_{i'\ell''}) = b_{i00}b_{i'01}b_{i'12}$$

$$E_1(t_{i\ell}t_{i'\ell'}t_{i\ell''}) = b_{i00}b_{i'01}b_{i12}$$

$$E_1(t_{i\ell}t_{i\ell'}t_{i'\ell''}t_{i''\ell'''}) = b_{i00}b_{i11}b_{i'02}b_{i''03}$$

$$E_1(t_{i\ell}t_{i'\ell}t_{i\ell'}t_{i'\ell''}) = b_{i00}b_{i'01}b_{i12}b_{i'13}$$

$$E_1(t_{i\ell}t_{i\ell'}t_{i\ell''}t_{i'\ell'''}) = b_{i00}b_{i11}b_{i22}b_{i'03}$$

$$E_1(t_{i\ell}t_{i'\ell'}t_{i''\ell''}t_{i'''\ell'''}) = b_{i00}b_{i01}b_{i''02}b_{i'''03}$$

Defining,

$$F_{1,1}(\underline{A}) = A_{42}$$

$$F_{1,2}(\underline{A}) = A_{43}A_{0-1} - A_{42}$$

$$F_{2,5}(\underline{A}) = A_{21}^2 - A_{42}$$

$$F_{3,1}(\underline{A}) = -2A_{41}$$

$$F_{3,2}(\underline{A}) = -2A_{43}A_{0-2} - 4A_{42}A_{0-1} + 6A_{41}$$

$$F_{3,3}(\underline{A}) = -2A_{43}A_{0-1}^2 + 2A_{43}A_{0-2} + 4A_{42}A_{0-1} - 4A_{41}$$

$$F_{4,5}(\underline{A}) = -2A_{21}A_{20} + 2A_{41}$$

$$F_{4,6}(\underline{A}) = -2A_{21}^2A_{0-1} + 4A_{21}A_{20} + 2A_{42}A_{0-1} - 4A_{41}$$

$$F_{5,5}(\underline{A}) = -4A_{31}A_{10} + 4A_{41}$$

$$F_{5,7}(\underline{A}) = -4A_{32}A_{10}A_{0-1} + 4A_{42}A_{0-1} + 4A_{31}A_{10} + 4A_{32}A_{1-1} - 8A_{41}$$

$$F_{6,11}(\underline{A}) = -2A_{21}A_{10}^2 + 2A_{21}A_{20} + 4A_{31}A_{10} - 4A_{41}$$

$$F_{7,1}(\underline{A}) = A_{40}$$

$$F_{7,2}(\underline{A}) = 3A_{41}A_{0-1} + 3A_{42}A_{0-2} + A_{43}A_{0-3} - 7A_{40}$$

$$F_{7,3}(\underline{A}) = 3A_{42}A_{0-1}^2 - 6A_{42}A_{0-2} - 9A_{41}A_{0-1} + 3A_{43}A_{0-2}A_{0-1} - 3A_{43}A_{0-3} + 12A_{40}$$

$$F_{7,4}(\underline{A}) = A_{43}A_{0-1}^2 - 3A_{43}A_{0-2}A_{0-1} + 2A_{43}A_{0-3} - 3A_{42}A_{0-1}^2 + 3A_{42}A_{0-2} + 6A_{41}A_{0-1} - 6A_{40}$$

$$F_{8,5}(\underline{A}) = 3A_{20}^2 - 3A_{40}$$

$$F_{8,6}(\underline{A}) = 3A_{20}A_{21}A_{0-1} - 3A_{20}^2 - 3A_{2-1}A_{21} - 3A_{41}A_{0-1} + 6A_{40}$$

$$F_{8,8}(\underline{A}) = 3A_{0-1}^2A_{21}^2 - 3A_{0-1}^2A_{42} - 3A_{0-2}A_{21}^2 + 3A_{0-2}A_{42} + 6A_{20}^2 - 12A_{20}A_{21}A_{0-1} + 12A_{2-1}A_{21} + 12A_{41}A_{0-1} - 18A_{40}$$

$$F_{8,10}(\underline{A}) = F_{8,6}(\underline{A})$$

$$F_{9,5}(\underline{A}) = 4A_{30}A_{10} - 4A_{40}$$

$$F_{9,7}(\underline{A}) = 8A_{31}A_{0-1}A_{10} - 8A_{31}A_{1-1} - 8A_{0-1}A_{41} + 4A_{0-2}A_{32}A_{10} - 4A_{0-2}A_{42} - 12A_{30}A_{10} - 4A_{32}A_{1-2} + 24A_{40}$$

$$F_{9,9}(\underline{A}) = 4A_{32}A_{10}A_{0-1}^2 - 4A_{42}A_{0-1}^2 - 4A_{32}A_{10}A_{0-2} + 4A_{42}A_{0-2} - 8A_{31}A_{10}A_{0-1} + 8A_{31}A_{1-1} + 8A_{30}A_{10} + 16A_{41}A_{0-1} - 24A_{40} - 8A_{32}A_{1-1}A_{0-1} + 8A_{1-2}A_{32}$$

$$F_{10,11}(\underline{A}) = 6A_{20}A_{10}^2 - 6A_{20}^2 - 12A_{30}A_{10} + 12A_{40}$$

$$F_{10,12}(\underline{A}) = 6A_{21}A_{10}^2A_{0-1} - 6A_{21}A_{20}A_{0-1} - 6A_{20}A_{10}^2 + 6A_{20}^2 + 12A_{1-1}A_{31} - 12A_{21}A_{1-1}A_{10} + 24A_{30}A_{10} - 12A_{31}A_{10}A_{0-1} + 12A_{21}A_{2-1} + 12A_{41}A_{0-1} - 36A_{40}$$

$$F_{11,13}(\underline{A}) = A_{10}^4 - 6A_{20}A_{10}^2 + 3A_{20}^2 + 8A_{30}A_{10} - 6A_{40}$$

$$F_{12,1}(\underline{A}) = A_{30}$$

$$F_{12,2}(\underline{A}) = A_{32}A_{0-2} + 2A_{31}A_{0-1} - 3A_{30}$$

$$F_{12,3}(\underline{A}) = A_{32}A_{0-1}^2 - A_{32}A_{0-2} - 2A_{31}A_{0-1} + 2A_{30}$$

$$F_{12,5}(\underline{A}) = 3A_{20}A_{10} - 3A_{30}$$

$$F_{12,6}(\underline{A}) = 3A_{21}A_{10}A_{0-1} - 3A_{20}A_{10} - 3A_{31}A_{0-1} - 3A_{21}A_{1-1} + 6A_{30}$$

$$F_{12,11}(\underline{A}) = A_{10}^3 - 3A_{20}A_{10} + 2A_{30}$$

If $F_{k,j}(\underline{A})$ ($k=1,\ldots 12$ and $j=1,\ldots,13$) is not given it is assumed to be zero, and

$$d_1 = \sum_{i=1}^{n} b_{i00}$$

$$d_2 = \sum_{i=1}^{n} b_{i00}b_{i11}$$

$$d_3 = \sum_{i=1}^{n} b_{i00}b_{i11}b_{i22}$$

$$d_4 = \sum_{i=1}^{n} b_{i00}b_{i11}b_{i22}b_{i33}$$

$$d_5 = \sum\sum_{i\neq i'}^{n} b_{i00}b_{i'01}$$

$$d_6 = \sum\sum_{i\neq i'}^{n} b_{i00}b_{i'01}b_{i'12}$$

$$d_7 = \sum\sum_{i\neq i'}^{n} b_{i00}b_{i11}b_{i'02}$$

$$d_8 = \sum\sum_{i\neq i'}^{n} b_{i00}b_{i12}b_{i'01}b_{i'03}$$

$$d_9 = \sum\sum_{i\neq i'}^{n} b_{i00}b_{i11}b_{i22}b_{i'03}$$

$$d_{10} = \sum\sum_{i\neq i'}^{n} b_{i00}b_{i12}b_{i'01}$$

$$d_{11} = \sum\sum\sum_{i\neq i'\neq i''}^{n} b_{i00}b_{i'01}b_{i''02}$$

$$d_{12} = \sum\sum\sum_{i\neq i'\neq i''}^{n} b_{i00}b_{i11}b_{i'02}b_{i''03}$$

$$d_{13} = \sum\sum\sum\sum_{i\neq i'\neq i''\neq i'''}^{n} b_{i00}b_{i'01}b_{i''02}b_{i'''03}$$

and

$$S_k = \sum_{j=1}^{13} F_{k,j}(A)d_j \quad k = 1,\ldots,12$$

we have

$$E_1D_1(\underline{Y},\underline{P},T) = \sum_{k=1}^{2} S_k$$

$$-2E_1D_2(\underline{Y},\underline{P},T) = \sum_{k=3}^{6} S_k$$

$$E_1D_3(\underline{Y},\underline{P},T) = \sum_{k=7}^{11} S_k$$

and

$$E_1v_4^2 = \sum_{k=1}^{11} S_k \ .$$

APPENDIX B

A derivation of Ev_6^2 for unequal probabilities and with replacement method for any n

From the characteristic function of the multinomial distribution we obtain the following needed moments:

$$Et_i = np_i$$

$$Et_i^2 = n(n-1)p_i^2 + np_i$$

$$Et_i t_{i'} = n(n-1)p_i p_{i'}$$

$$Et_i t_{i'} t_{i''} = np_i(n-1)p_{i'}(n-2)p_{i''}$$

$$Et_i^2 t_{i'} = n(n-1)(n-2)p_i^2 p_{i'} + n(n-1)p_i p_{i'}$$

$$Et_i^4 = n(n-1)(n-2)(n-3)p_i^4 + 6n(n-1)(n-2)p_i^3 + 7n(n-1)p_i^2 + np_i$$

$$Et_i^2 t_{i'}^2 = n(n-1)(n-2)(n-3)p_i^2 p_{i'}^2 + n(n-1)(n-2)p_i^2 p_{i'} + n(n-1)(n-2)p_i p_{i'}^2 + n(n-1)p_i p_{i'}$$

$$Et_i^3 t_{i'} = n(n-1)(n-2)(n-3)p_i^3 p_{i'} + 3n(n-1)(n-2)p_i^2 p_{i'} + n(n-1)p_i p_{i'}$$

$$Et_i^2 t_{i'} t_{i''} = n(n-1)(n-2)(n-3)p_i^2 p_{i'} p_{i''} + n(n-1)(n-2)p_i p_{i'} p_{i''}$$

$$Et_i t_{i'} t_{i''} t_{i'''} = n(n-1)(n-2)(n-3)p_i p_{i'} p_{i''} p_{i'''}$$

Writing v_6 as

$$v_6 = 1/n(n-1)\left[\Sigma t_i y_i^2/p_i^2 - n\hat{Y}_6^2\right]$$

where

$$\hat{Y}_6 = \Sigma t_i y_i/np_i$$

$$t_i = \begin{cases} 1 \text{ if i-th unit is in the sample} \\ 0 \text{ otherwise} \end{cases}$$

using the above moments of the multinomial distribution we have,

$$Ev_6^2 = \left[1/n(n-1)\right]^2 E\left[\left(\sum_{i=1}^{N} t_i y_i^2/p_i^2\right)^2 - 2n\hat{Y}_6^2 \sum_{i=1}^{N} t_i y_i^2/p_i^2 + n^2\hat{Y}_6^4\right]$$

where

$$E\left(\Sigma t_i y_i^2/p_i^2\right)^2 = nA_{43} + n(n-1)A_{21}^2$$

$$n^2E(\hat{Y}_6^2\,\Sigma t_i y_i^2/p_i^2) = nA_{43} + n(n-1)A_{21}^2 + 2n(n-1)A_{10}A_{32} + n(n-1)(n-2)A_{21}A_{10}^2$$

$$n^4E\hat{Y}_6^4 = n(n-1)(n-2)(n-3)A_{10}^4 + 6n(n-1)(n-2)A_{10}^2A_{21} + 4n(n-1)A_{32}A_{10} + 3n(n-1)A_{21}^2 + nA_{43}$$

$$A_{ij} = \sum_{t=1}^{N} y_t^i/p_t^j$$

APPENDIX C

A derivation of ϵEv_4^2 for the R.H.C. method for any n.

Substituting $y_i = \beta x_i + e_i$ into Ev_4^2, which involves replacing each y_i by e_i since it is since it is easily shown that Ev_4^2 is independent of the βx_i term, we have after taking expectation,

$$\epsilon Ev_4^2 = \sum_{k=1}^{8} R_k$$

where

$$R_k = \sum_{j=1}^{13} H_{k,j}(B)d_j \qquad k = 1, \ldots, 8 .$$

Where the d_j's are given in Appendix A and

$$H_{1,1}(\underline{B}) = B_{42}$$

$$H_{1,2}(\underline{B}) = B_{43}\, B_{0-1} - B_{42}$$

$$H_{2,5}(\underline{B}) = B_{21}^2 - (1/3)B_{42}$$

$$H_{3,1}(\underline{B}) = -2B_{41}$$

$$H_{3,2}(\underline{B}) = -2B_{43}\, B_{0-2} - 4B_{42}B_{0-1} + 6B_{41}$$

$$H_{3,3}(\underline{B}) = -2B_{43}\, B_{0-1}^2 + 2B_{43}\, B_{0-2} + 4B_{42}\, B_{0-1} - 4B_{41}$$

$$H_{4,5}(\underline{B}) = -2B_{21}\, B_{20} + (2/3)B_{41}$$

$$H_{4,6}(\underline{B}) = (2/3)B_{42}\, B_{0-1} - 2B_{0-1}\, B_{21}^2 - (4/3)B_{41} + 4B_{21}\, B_{20}$$

$$H_{7,1}(\underline{B}) = B_{40}$$

$$H_{7,2}(\underline{B}) = 3B_{41}\, B_{0-1} + 3B_{42}\, B_{0-2} + B_{43}\, B_{0-3} - 7B_{40}$$

$$H_{7,3}(\underline{B}) = 3B_{42}\, B_{0-1}^2 - 6B_{42}\, B_{0-2} - 9B_{41}\, B_{0-1} + 3B_{43}\, B_{0-2}\, B_{0-1} - 3B_{43}\, B_{0-3} + 12B_{40}$$

$$H_{7,4}(\underline{B}) = B_{43}\, B_{0-1}^2 - 3B_{43}\, B_{0-2}\, B_{0-1} + 2B_{43}\, B_{0-3} - 3B_{42}\, B_{0-1}^2 + 3B_{42}\, B_{0-2} + 6B_{41}\, B_{0-1} - 6B_{40}$$

$$H_{8,5}(\underline{B}) = 3B_{20}^2 - B_{40}$$

$$H_{8,6}(\underline{B}) = 3B_{0-1}\, B_{21}\, B_{20} + 2B_{40} - 3B_{20}^2 - 3B_{21}\, B_{2-1} - B_{41}\, B_{0-1}$$

$$H_{8,8}(\underline{B}) = 3B_{0-1}^2\, B_{21}^2 - B_{0-1}^2\, B_{42} - 3B_{0-2}\, B_{21}^2 + B_{0-2}\, B_{42}$$

$$+ 6B_{20}^2 - 12B_{0-1}\, B_{21}\, B_{20} + 12B_{21}\, B_{2-1} + 4B_{41}\, B_{0-1} - 6B_{40}$$

$$H_{8,10}(\underline{B}) = H_{8,6}(\underline{B})$$

Where if $H_{k,j}(\underline{B})$ ($k = 1, \ldots 8$ and $j = 1, \ldots 13$) is not given it is zero, and

$$B_{0j} = \sum_{t=1}^{N} 1/p_t^j$$

$$B_{2j} = \sum_{t=1}^{N} x_t^g/p_t^j$$

$$B_{4j} = \sum_{t=1}^{N} 3x_t^{2g}/p_t^j$$

APPENDIX D

A scheme to calculate the conditional probabilities $p(s|i)$ and $p(s|ii')$ for Murthy's method

Scheme:

1. Rearrange the integers of a given sample s, $\ell_1 \ldots \ell_n$, $1 \le \ell_i \le N$ such that $\ell_1 < \ell_2 < \ldots \ell_n$.
2. Form n groups of the n! permutations by placing ℓ_i as the first element of the $(n-1)!$ permutations in group i.
3. Within group i, i = 1, ..., n, form n-1 subgroups by placing $\ell_{i'}$, $i' \neq i$, as the second element of the $(n-2)!$ permutations in subgroup i' with the elements with subgroups arranged by the ascending order of the ℓ_i's.

Hence, by definition

$$p\{s|i\} = \sum_{\text{Group } i} p_{3\ell_1\ldots\ell_n}$$

$$p\{s|ii'\} = \sum_{\text{Group } i(i')} p_{3\ell_1\ldots\ell_n} + \sum_{\text{Group } i'(i)} p_{3\ell_1\ldots\ell_n}$$

where

$$p_{3\ell_1\ldots\ell_n} = p_{\ell_1}\left[p_{\ell_2}/(1-p_{\ell_1}) \ldots p_{\ell_n}/(1-p_{\ell_1}-\ldots-p_{\ell_{n-1}})\right].$$

Diagram of Scheme

Group	Sub-group	Permutation	
	Sub-group 1(2)	$\ell_1\ell_2\ldots$ $\ldots$ $\ell_1\ell_2\ldots$	(n-2)! permutations
	...	...	
Group (1)	Sub-group 1(i')	$\ell_1\ell_{i'}\ldots$ $\ldots$ $\ell_1\ell_{i'}\ldots$	(n-2)! permutations
	...		
	Sub-group 1(n)	$\ell_1\ell_n\ldots$ $\ldots$ $\ell_1\ell_n$	(n-2)! permutations
...	...		...
	Sub-group i(1)	$\ell_i\ell_1\ldots$ $\ldots$ $\ell_i\ell_1\ldots$	(n-2)! permutations
	...	...	
Group (i)	Sub-group i(i')	$\ell_i\ell_{i'}\ldots$ $\ldots$ $\ell_i\ell_{i'}\ldots$	(n-2)! permutations
	...	...	
	Sub-group i(n)	$\ell_i\ell_n\ldots$ $\ldots$ $\ell_i\ell_n\ldots$	(n-2)! permutations
...	...	...	...
	Sub-group n(1)	$\ell_n\ell_1\ldots$ $\ldots$ $\ell_n\ell_1\ldots$	(n-2)! permutations
	...	...	
Group (n)	Sub-group n(i')	$\ell_n\ell_{i'}\ldots$ $\ldots$ $\ell_n\ell_{i'}\ldots$	(n-2)! permutations
	...	...	
	Sub-group n(n-1)	$\ell_n\ell_{n-1}\ldots$ $\ldots$ $\ell_n\ell_{n-1}\ldots$	(n-2)! permutations

APPENDIX E

Derivations of $\epsilon v_2'(s')$ and $\epsilon v_2'^2(s')$ for Des Raj's method

Writing Des Raj's variance estimator as

$$n^2(n-1)\, v_2(s') = \sum_{\substack{i<i' \\ s\supset ii'}}^{n} \sum (t_i - t_{i'})^2$$

where s' denotes one of n! possible ordering of a given sample s and

$$t_1 = y_1/p_1$$

$$\cdots$$

$$t_i = \sum_{r=1}^{i-1} y_r + \left(1 - \sum_{r=1}^{i-1} p_r\right) y_i/p_i$$

$$\cdots$$

$$t_n = \sum_{r=1}^{n-1} y_r + \left(1 - \sum_{r=1}^{n-1} p_r\right) y_n/p_n$$

The 'trick' of this derivation is to write $(t_i - t_{i'})$ for a given s' as

$$(t_i - t_{i'}) = \sum_{j=1}^{n} C_{ii'j} y_j$$

where for $i < i'$

$$C_{ii'j} = \begin{cases} 0 & j < i \\ \left(\sum_{r=1}^{i} p_r - 1\right)/p_i & j = i \\ 1 & i < j < i' \\ \left(1 - \sum_{r=1}^{i-1} p_r\right)/p_i & j = i' \\ 0 & j > i' \end{cases}$$

and for $i > i'$ we have

$$C_{iij} = \begin{cases} 0 & j < i' \\ -C_{ii'i} & j = i' \\ -1 & i' < j < i \\ -C_{ii'i'} & j = i \\ 0 & j > i \end{cases}$$

Now, by letting $v_2'(s')$ be $v_2(s')$ under the assumption of the super population model we have

$$n^2(n-1) v_2'(s') = \sum_{\substack{i<i' \\ s'\supset ii'}}^{n} \sum \left[\sum_{j=1}^{n} C_{ii'j}\, e_j\right]^2$$

and

$$2n^2(n-1)\epsilon v_2'(s') = \sum_{\substack{i\neq i' \\ s'\supset ii'}}^{n} \sum \left[\sum_{j=1}^{n} C_{ii'j}^2\, x_j^g\right]$$

Similarly, $\epsilon v_2'^2(s')$ is obtained after considerable manipulation as

$$4n^4(n-1)^2\, \epsilon v_2'^2(s') = (2n^2(n-1)\, \epsilon v_2'(s')^2$$

$$+ 4 \sum_{\substack{i\neq i' \\ s\supset ii'}}^{n} \sum \left[\sum_{j=1}^{n} C_{ii'j}^2\, x_j^g\right]^2$$

$$+ 4 \sum_{\substack{i\neq i'\neq i'' \\ s\supset ii'i''}} \sum^{n} \sum \left[\sum_{j=1}^{n} C_{ii'j} C_{ii''j}\, x_j^g\right]^2$$

$$+ 4 \sum_{\substack{i\neq i'\neq i'' \\ s'\supset i\, i'i''}} \sum^{n} \sum \left[\sum_{j=1}^{n} C_{ii'j} C_{i'i''j}\, x_j^g\right]^2$$

$$+ 2 \sum_{\substack{i\neq i'\neq i''\neq i''' \\ s\supset ii'i''i'''}} \sum^{n} \sum \sum \left[\sum_{j=1}^{n} C_{ii'j}\, C_{i''i'''j}\, x_j^g\right]^2 .$$

TABLE 3.1 Description of the natural population for n = 3

Pop. No.	Source	y	x	N	C.V.(y)	C.V.(x)	ρ
1	Horvitz Thompson (1952)	No. of households	Eye-estimated no. of households	20	0.44	0.40	.87
2	Des Raj (1965)	No. of households	Eye-estimated no. of households	20	0.44	0.41	.66
3	Rao (1963)	Corn acreage in 1960	Corn acreage in 1958	14	0.39	0.43	.93
4	Kish (1965)	No. of rented dwellings	Total no. of dwellings	15	1.37	1.06	.98
5	Cochran (1963)	Wt. of peaches	Eye-estimated wt. of peaches	10	0.19	0.17	.97
6	Hanurav (1967)	Population in 1960	Population in 1950	16	0.66	0.65	.99
7	Cochran (1963)	No. of persons per block	No. of rooms per block	10	0.15	0.14	.65
8	Cochran (1963)	No. of people in 1930	No. of people in 1920	20	0.85	0.93	.97
9	Cochran (1963)	No. of people in 1930	No. of people in 1920	20	0.71	0.82	.95
10	Sukhatme (1954)	No. of wheat A's in 1937	No. of wheat A's in 1936	20	0.76	0.74	.99
11	Sampford (1962)	Oat A's in 1957	Total A's in 1947	20	0.62	0.70	.83
12	Sukhatme (1954)	Wheat A's	No. of villages	20	0.59	0.51	.52
13	Yates (1960)	Volume of timber	Eye-estimate of volume	20	0.52	0.48	,50
14	Yates (1960)	No. of absentees	Total no. of persons	20	0.53	0.46	.67

TABLE 3.2. Percent gains in efficiency of the estimators over Sampford's estimator for $n = 3$.

Pop. No.	Carroll-Hartley	Fellegi	Murthy	Des Raj	R.H.C.	Lahiri	With Rep.
1	+0	-0	-2	-3	-5	-22	-15
2	-0	-1	-0	-1	-3	-14	-13
3	-0	-2	3	1	+0	3	-14
4	-1	-8	1	-4	-3	-34	-24
5	-0	-1	1	-3	-2	2	-22
6	+0	-2	-0	-3	-7	-25	-19
7	+0	-1	1	-2	-1	6	-21
8	-0	-5	8	6	3	13	- 7
9	-0	-4	9	8	10	563	- 1
10	+0	-2	2	+0	-3	-10	-13
11	-0	-2	1	-0	-3	-16	-13
12	-0	-2	4	3	4	44	- 7
13	-0	-1	1	-1	-2	- 9	-12
14	-0	-1	2	1	1	10	- 9

TABLE 3.3. Percent gains in efficiency of the variance estimators over Sampford's variance estimator for $n = 3$.

Pop. No.	Carroll-Hartley	Fellegi	Murthy	Des Raj	R.H.C.	Lahiri	With Rep.	Modified Lahiri
1	+0	1	-5	-6	-10	-100	-18	- 99
2	+0	-0	2	1	2	- 95	- 6	- 88
3	-0	-1	13	10	19	-100	- 7	-100
4	1	-0	45	32	47	- 99	22	- 99
5	-0	-1	6	4	9	-100	-17	-100
6	-0	-2	8	6	5	-100	- 9	-100
7	+0	-1	6	4	10	-100	-15	-100
8	+0	+0	15	11	21	- 99	2	- 99
9	+0	+0	12	10	22	- 34	4	71
10	+0	-2	22	22	42	-100	20	-100
11	1	-0	19	19	32	- 99	14	- 97
12	+0	-1	9	9	18	- 78	4	- 14
13	-0	-2	10	10	19	- 93	4	- 82
14	+0	-0	7	6	12	- 96	- 1	- 91

TABLE 3.4. Percent gains in average efficiency of the estimators over Sampford's estimator (under the super population model for $g = 1.5$, 1.75, and 2.00) for $n = 3$.

(Natural Populations)

	$g = 1.50$			$g = 1.75$			$g = 2.00$		
Pop. No.	Murthy	Des Raj	R.H.C.	Murthy	Des Raj	R.H.C.	Murthy	Des Raj	R.H.C.
1	+0	-1	-2	+0	-1	-2	-0	-1	-2
2	+0	-1	-2	+0	-1	-2	-0	-1	-3
3	1	-2	-3	+0	-2	-4	-0	-3	-5
4	4	-2	-11	1	-6	-16	-2	-10	-20
5	+0	-4	-3	+0	-4	-3	-0	-4	-4
6	2	-1	-4	+0	-2	-6	-1	-3	-8
7	+0	-4	-3	+0	-4	-3	-0	-4	-3
8	3	+0	-5	1	-2	-8	-2	-5	-11
9	2	-0	-4	1	-2	-7	-1	-3	-9
10	1	-0	-4	1	-1	-6	-0	-2	-7
11	1	-0	-3	1	-1	-5	-0	-2	-6
12	1	-1	-2	+0	-1	-3	-0	-1	-4
13	1	-1	-2	+0	-1	-3	-0	-1	-3
14	1	-1	-2	+0	-1	-2	-0	-1	-3

TABLE 3.5. Percent gains in average efficiency of the variance estimators over Sampford's variance estimator (under the assumption of a super population model for $g = 1.50, 1.75$, and 2.00) for $n = 3$.

Natural Populations

	g = 1.50					g = 1.75					g = 2.00				
Pop. No.	Mur.	Des Raj	R.H.C.	Carroll Hartley	Fellegi	Mur.	Des Raj	R.H.C.	Carroll Hartley	Fellegi	Mur.	Des Raj	R.H.C.	Carroll Hartley	Fellegi
1	4	3	6	+0	-4	3	2	3	+0	-4	1	1	1	+0	-3
2	4	4	6	+0	-4	3	2	4	+0	-4	2	1	1	+0	-4
3	9	8	12	+0	-7	7	5	8	+0	-7	4	3	3	+0	-6
4	56	53	90	1	-16	50	45	62	1	-18	39	32	35	+0	-18
5	3	1	3	+0	-4	2	-0	2	+0	-4	1	-1	+0	+0	-4
6	17	16	25	+0	-10	14	12	17	+0	-10	10	7	9	+0	-9
7	2	-0	1	+0	-3	1	-1	1	+0	-3	1	-1	-0	-0	-3
8	31	28	44	+0	-14	29	26	37	+0	-15	25	20	24	+0	-15
9	20	19	30	+0	-10	17	15	21	+0	-10	12	9	10	+0	-10
10	13	13	21	+0	-7	9	7	10	+0	-7	4	3	2	+0	-6
11	13	12	20	+0	-8	10	9	12	+0	-7	6	4	4	+0	-7
12	6	6	9	+0	-5	4	3	5	+0	-5	2	1	1	+0	-4
13	6	5	8	+0	-5	4	3	4	+0	-4	2	1	1	+0	-4
14	5	5	8	+0	-5	4	3	4	+0	-4	2	1	1	+0	-4

TABLE 4.1. Description of the natural population for $n = 4$.

Pop. No.	Source	y	x	N	C.V.(y)	C.V.(x)	ρ
1	Horvitz & Thompson (1952)	No. of households	Eye-estimated no. of households	16	.40	.43	.91
2	Rao (1963)	Corn acreage in 1960	Corn acreage in 1958	14	.39	.43	.93
3	Kish (1965)	No. of rented dwelling units	Total no. of dwellings	15	1.37	1.06	.98
4	Cochran (1963)	Wt. of peaches	Eye-estimated wt. of peaches	10	.19	.17	.97
5	Hanurav (1967)	Population in 1960	Population in 1950	16	.66	.65	.99
6	Cochran (1963)	No. of persons per block	No. of rooms per block	10	.15	.14	.65
7	Cochran (1963)	No. of people in 1930	No. of people in 1920	12	.78	.95	.96
8	Sukhatme (1954)	No. of wheat A's in 1937	No. of wheat A's in 1936	13	.80	.76	.98
9	Sampford (1962)	Oats A's in 1957	Total A's in 1947	14	.65	.69	.75
10	Yates (1960)	Volume of timber	Eye-estimated volume of timber	11	.37	.45	.72

TABLE 4.2. Percent gains in efficiency of the estimators over Sampford's estimator for n = 4.

Natural Populations						
Pop. No.	Carroll-Hartley	Murthy	Des Raj	R.H.C.	Lahiri	With Repl.
1	-0	-0	- 4	- 5	-16	-24
2	-0	4	+ 0	- 1	6	-21
3	-1	-4	-14	-24	-44	-39
4	-0	2	- 6	- 5	3	-32
5	+0	-1	- 6	-10	-31	-28
6	+0	2	- 5	- 3	8	-31
7	-0	33	25	33	849	- 3
8	+0	+0	-10	-18	-34	-37
9	+0	-2	- 9	-15	-30	-33
10	-0	4	- 4	- 4	- 0	-30

TABLE 4.3. Percent gains in efficiency of the variance estimators over Sampford's variance estimator for n = 4 .

Pop. No.	Carroll-Hartley	Murthy	Des Raj	R.H.C.	Lahiri	With Repl.	Modified Lahiri
1	+0	4	+0	2	-100	-16	-100
2	-0	19	12	27	-100	-10	-100
3	+1	83	56	140	- 99	52	- 99
4	-0	10	3	12	-100	-23	-100
5	-0	13	8	8	-100	-12	-100
6	-0	11	4	15	-100	-21	-100
7	+1	43	23	75	- 75	13	- 33
8	+2	120	128	242	-100	135	-100
9	+0	96	100	153	- 98	92	- 96
10	-0	21	13	31	- 99	- 5	- 99

TABLE 4.4 Description of the natural populations for n = 4 under the super population model.

Natural Populations									
	g = 1.50			g = 1.75			g = 2.00		
Pop. No.	Murthy	Des Raj	R.H.C.	Murthy	Des Raj	R.H.C.	Murthy	Des Raj	R.H.C.
1	1	- 9	-10	+0	-10	-11	-1	-11	-13
2	1	- 8	- 8	-0	- 9	- 9	-1	-10	-11
3	6	-15	-25	-2	-22	-33	-9	-30	-40
4	+0	- 8	- 7	-0	- 8	- 7	-0	- 8	- 8
5	4	-12	-19	-2	-18	-26	-8	-24	-32
6	+0	- 8	- 7	+0	- 8	- 7	-0	- 8	- 7
7	1	-10	-13	-0	-11	-15	-1	-13	-17
8	4	-13	-21	-1	-17	-26	-4	-21	-30
9	3	-11	-16	-1	-15	-21	-6	-20	-26
10	2	- 9	-12	-0	-11	-14	-2	-13	-17

TABLE 4.5 Percent gains in average efficiency of the estimators over Sampford's estimator (using the super population model for g = 1.50, 1.75, and 2.00) for n = 4.

Pop. No.	Source	y	x	N	C.V.(y)	C.V.(x)	ρ
1	Horvitz & Thomp. (1952)	No. of households	Eye-estimated no. of households	10	.41	.37	.89
2	Rao (1963)	Corn A's in 1960	Corn A's in 1958	10	.23	.30	.81
3	Kish (1965)	No. of rented dwelling units	Total no. of dwelling units	10*	1.41	.82	.93
4	Cochran (1963)	Wt. of peaches	Eye-estimated wt. of peaches	10	.19	.17	.97
5	Hanurav (1967)	Population in 1960	Population in 1950	10	.75	.73	.99
6	Cochran (1963)	No. of persons per block	No. of rooms per block	10	.15	.14	.65
7	Cochran (1963)	No. of people in 1930	No. of people in 1920	10	.31	.47	.34
8	Suk. (1954)	No. of wheat A's in 1937	No. of wheat A's in 1936	10	.75	.69	.98
9	Sampford (1962)	Oats A's in 1957	Total A's in 1947	10	.58	.64	.91
10	Yates (1960)	Volume of timber	Eye-estimated vol. of timber	10	.39	.47	.72

* one x-value changed so that $\pi_i \leq nP_i$ for i = 1,...,10

TABLE 4.6. Percent gains in average efficiency of the variance estimators over Sampford's variance estimator (using the super population model for g = 1.50, 1.75, and 2.00) for n = 4.

Natural Populations												
	g = 1.50				g = 1.75				g = 2.00			
Pop. No.	Carroll Hartley	Murthy	Des Raj	R.H.C.	Carroll Hartley	Murthy	Des Raj	R.H.C.	Carroll Hartley	Murthy	Des Raj	R.H.C.
1	+0	23	17	28	+0	19	12	19	+0	14	7	10
2	+0	15	10	17	+0	13	7	12	+0	10	4	7
3	3	164	159	257	2	163	151	201	1	149	129	145
4	+0	5	-0	3	+0	4	-1	2	+0	3	-2	+0
5	+0	111	101	158	+0	112	98	134	-0	106	88	103
6	+0	3	-2	1	+0	2	-3	+0	+0	2	-3	-1
7	+0	34	29	45	+0	23	15	22	+0	17	9	11
8	2	90	85	135	1	75	64	84	1	60	47	52
9	+0	74	64	101	+0	73	61	85	+0	68	54	65
10	+0	40	33	50	+0	35	27	38	+0	28	20	25

AN APPLICATION OF THE ANALYSIS OF VARIANCE TO LARGE SAMPLE SURVEYS

Murle J. Atherton, National Center for Health Statistics

1. Introduction

A sample survey of any magnitude usually involves at least one stage of cluster sampling. Most often inherent to this type of design are correlated observations. These correlations are dependent, for example, on how the data is classified. If, say, only cluster means or totals, or some combination of them, are to be used as observations, then the correlations will be zero. However, if units from the same cluster are contained in different observations then non-zero correlations will probably exist. Also inherent to large surveys is the problem of inequality of variance--many times extreme inequality. Surveys such as those conducted by the National Center for Health Statistics and the Bureau of the Census yield data of this type.

This paper treats a common special case where the data in question fits a two-way classification model with one observation per cell. An extension to more than two variables would be theoretically no more difficult.

2. Statistical Model

The "usual" two-way classification model with one observation per cell is of the form Γ, where the error terms are independently and identically distributed,

$$\Gamma: \begin{cases} y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \\ \epsilon_{ij} \overset{iid}{\rightarrow} N(0, \sigma^2) \\ \Sigma\alpha_i = \Sigma\beta_j = 0 \end{cases}$$

but the model we shall consider is of the form Γ', where the error terms are not all independent and do not all have the same variance.

$$\Gamma': \begin{cases} y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \\ \epsilon_{ij} \overset{d}{\rightarrow} N(0, \sigma^2_{ij}) \\ \sigma_{ij,\, i'j'} \neq 0 \text{ for some } (i, j) \neq (i', j') \\ \Sigma\alpha_i = \Sigma\beta_j = 0 \end{cases}$$

or in matrix notation

$$\Gamma': \begin{cases} Y = X'\beta + \epsilon \\ \epsilon \overset{d}{\rightarrow} N(0, \Sigma) \\ \Sigma\alpha_i = \Sigma\beta_j = 0 \end{cases}$$

3. Analysis

By appealing to the asymptotic properties of estimators in large samples it seems reasonable to assume we can make "good" estimates of the covariance matrix Σ. Or that at least we can make good estimates of the ratios of the components of Σ; i.e., estimate C, where $\Sigma = \sigma^2 C$.

There are pros and cons for each of these two estimates. If we assume Σ is "known," then we have one degree of freedom to test for interactions, which is, of course, desirable. However, this test is generally very sensitive to any error in the estimate, $\hat{\Sigma}$, of Σ. For example, any bias in $\hat{\Sigma}$, is directly reflected in the mean square (MS) used to test for interactions. Also the magnitude of the MS, which is the χ^2-statistic of interest, is heavily dependent on the magnitude of the observations. Therefore, a very small relative difference in the observations may result in a large MS.

On the other hand, if we assume that only the ratio matrix C is known and that there are no interactions and use an F-test, then any bias is immediately removed by the F-ratio. However, if significant interactions do exist then any inferences we may make concerning the main effects may not be correct.

A very practical solution to this problem would be to start the analysis under the model Γ''', where Γ''' differs from Γ' through the addition of a set of interaction terms $\{\gamma_{ij}\}$.

$$\Gamma''': \begin{cases} y_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ij} \\ \epsilon_{ij} \overset{d}{\rightarrow} N(0, \sigma^2_{ij}) \\ \sigma_{ij,\, i'j'} \neq 0 \text{ for some } (i, j) \neq (i', j') \\ \Sigma\alpha_i = \Sigma\beta_j = \sum_i \gamma_{ij} = \sum_j \gamma_{ij} = 0 \end{cases}$$

Then assume we know Σ and use a χ^2- test for interactions. Under these assumptions one could also test main effects using the χ^2-test. But for reasons mentioned previously this is not done. After making this one test abandon Γ''' and assume Γ'. Now assume C is known and use F-tests to make inferences about the main effects. This procedure will give a test for interactions (which should not be interpreted too literally) and hence, a better insight into the validity of tests on main effects which assume an additive model.

Two additional qualifications should be made concerning the test for interactions. First, the MS used as the χ^2- statistic in this test not only contains interaction effects but also the random error component of the model.

And secondly, when significant interactions exist one should consider main effects to be "different" even if tests involving these effects are not significant. The reasoning being that when the effects of the levels of one factor are averaged over the levels of the other, no difference of these "averaged" effects has been shown.

Now to consider the working theory of the procedure, let us continue by assuming estimates are made for the elements of Σ. We know from matrix algebra[1] that there exists a non-singular matrix P, such that $P'\Sigma P = I$. Then if we transform the observations and the design matrix according to P; i.e., let

$$Z = P'Y$$

$$W' = P' X'$$

the vector Z satisfies

$$\Gamma'': \begin{cases} Z = W'\beta + \epsilon \\ \epsilon \overset{d}{\to} N(0, I) \\ \Sigma\alpha_i = \Sigma\beta_j = 0 \end{cases}$$

I comment here that the $\hat{\beta}$ obtained under Γ'' have the same expected value[2] as those obtained under Γ'; so no change in notation is made.

To estimate β we now need to minimize the following SS with respect to β:

$$S = (Z - W'\beta)' (Z - W'\beta)$$

$$\neq \sum_{ij} (Z_{ij} - \mu - \alpha_i - \beta_j)^2$$

or

$$S = (Y - X'\beta)' \Sigma^{-1} (Y - X'\beta)$$

And since this minimum is the L.S. estimate we need only solve the modified normal equations[2]

$$\hat{\beta} = (X\Sigma^{-1}X' + HH')^{-1} X\Sigma^{-1} Y$$

where the matrix H incorporates into the solution the side conditions $\Sigma\alpha_i = \Sigma\beta_j = 0$; i.e., $H'\beta = 0$.

Similarly for the row and column hypotheses we have the solutions

$$\hat{\beta}_{H_R} = (X_{H_R}\Sigma^{-1}X'_{H_R} + H_{H_R} H'_{H_R})^{-1} X_{H_R}\Sigma^{-1} Y$$

$$\hat{\beta}_{H_C} = (X_{H_C}\Sigma^{-1}X'_{H_C} + H_{H_C} H'_{H_C})^{-1} X_{H_C}\Sigma^{-1} Y$$

where the matrices X_{H_R}, X_{H_C}, H_{H_R}, and H_{H_C} are obtained by deleting the rows of X and H that correspond to the hypothesis in question.

The SS's needed for the tests of hypothesis are found by first substituting these parameter estimates into the equations

$$S = (Y - X'\beta)'\Sigma^{-1}(Y - X'\beta)$$

$$S_R = (Y - X'_{H_R}\beta_{H_R})'\Sigma^{-1}(Y - X'_{H_R}\beta_{H_R})$$

$$S_C = (Y - X'_{H_C}\beta_{H_C})'\Sigma^{-1}(Y - X'_{H_C}\beta_{H_C})$$

and we obtain

$$SS_{Row} = S_R - S \neq \sum_i \hat{\alpha}_i^2$$

$$SS_{Col} = S_C - S \neq \sum_j \hat{\beta}_j^2$$

$$SS_e = S$$

The above inequalities are pointed out because under Γ these inequalities become equalities. This follows from the fact that even though Γ and Γ'' appear to be the same, the design matrix under Γ'' is such that the estimates of $\{\alpha_i\}$ and $\{\beta_j\}$ under the hypotheses are not necessarily the same as when no hypotheses are imposed. Or geometrically we say that $\{\alpha_i\}$ and $\{\beta_j\}$ are not necessarily from orthogonal spaces.

It is possible that the SS_{Row} and the SS_{Col} might be negative for several reasons. (1) One of the matrices to be inverted may be "nearly" singular or "ill-conditioned." (2) Due to an error in estimation $\hat{\Sigma}$ may not be positive definite. (3) Cumulative round-off error in the analytical program may introduce negative SS's.

4. Example

Let us now consider an example with observations and covariance matrix given on next page. Note that $y_1 = 23.35$, $y_2 = 1.13$, ..., $y_9 = 3.08$, $\sigma^2_{y_1} = 129.57$, $\sigma^2_{y_2} = 79.34$, ..., $\sigma^2_{y_9} = 156.07$.

Using the scheme developed above the ANOVA table obtained from these observations is then

Source	SS	DF	MS	F-Ratio
Row	47.27654	2	23.63826	15.92203
Column	9.85925	2	4.92963	3.32045
Error	5.93851	4	1.48463	

Suppose we wish to test at the 95% level. The first step in the analysis is then to compare the mean square for error with the tabled χ^2-value with 4 d.f. and 95% confidence; i.e., $\chi^2_{.05;\,4} = 9.48773$ for a one-sided test, or $\chi^2_{.975;\,4} = .484419$ and $\chi^2_{.025;\,4} = 11.1433$ for a two-sided test.

Thus, in either case, we accept the hypothesis that there are no significant interactions. Next we compare the F-ratios corresponding to the row and column hypotheses to the tabled value $F_{.05;\,2,\,4} = 6.9443$.

We reject the hypothesis concerning row effects; i.e., we conclude there is a significant difference between rows.

To see what effect small errors in the estimation of Σ have on the analysis let us hold the variances constant and change the covariances in four ways. The results are given on the following 3 pages, with the Σ_i's denoting changes in the original covariance matrix.

If we assume independence and equal variances the ANOVA table is

Source	SS	DF	MS	F-Ratio
Row	121.38843	2	60.69421	0.52251
Column	27.98560	2	13.99280	0.12046
Error	464.63818	4	116.15955	

And we see that we change the conclusion regarding row differences to "not significant."

Small errors in Σ change the F-value for rows from 1.2% under Σ_1 to 13.2% under Σ_4. For the column F-values the changes range from 2.8% under Σ_4 to 8.3% under Σ_1. However, these respective changes are not nearly so severe as the 96.7% for rows and 96.4% for columns encountered when we assume independence and equality of variance.

5. Summary of Emperical Work

To summarize the emperical work that was done look at the set of detailed tables. Table 1 relates the sizes of the tables analyzed, the number of sets of data for each size, and the number of conclusions that were changed. By "number of conclusions changed" is meant the number of times the results of the original tests of hypotheses were changed, first when an error was made in Σ, and second when independence and equal variance was assumed. For the 82 2 x 2 tables analyzed 1.2% of the row hypotheses and 1.2% of the column hypotheses were wrong when small errors were made in Σ as opposed to 11% and 18.3% when independence and equal variance was assumed. For the 101 3 x 3 tables these respective percentages were 12% and 3% as opposed to 50% and 12%, for the 4 x 3 tables 0.0% and 2% as opposed to 20% and 30%, for the 5 x 2 tables 4.8% and 0.0% as opposed to 19% and 10%, and for the 5 x 4 tables 1.7% and 1.7% as opposed to 27% and 9%. Table 2 relates the average percent change in the F-statistics for the same data from which Table 1 was obtained. In interpreting Table 2 if we would keep in mind that in standard F-tables a change of significance level from say 95% to 90% requires a change of about 25% in the tabled value, then these results might be more meaningful. From these two tables we see that though errors in Σ are not too serious, the assumption of independence and equal variance can lead to quite unreliable test statistics.

For practical purposes it was hoped that small correlations (less than 10%) could be ignored. Emperically it was found that if variances are equal, little "harm" comes from ignoring small correlations and proceeding classically. For example, in Table 3 we see that for the 120 3 x 3 tables that were analyzed by assuming correlations of less than 10% to be zero, only 4.1% of the conclusions reached were incorrect for the row hypotheses and 1.7% were incorrect for the column hypotheses. For the 40 4 x 3 tables analyzed these percentages were 2.5% and 0.0%, and for the 80 5 x 4 tables analyzed there were no incorrect conclusions.

However, if variances are not equal (differences ranging from 2 to 100 times one another) and we proceed under the assumption of independence and equal variance, then a great deal of accuracy is lost as is seen in Table 4. For example, for the 124 3 x 3 tables analyzed, 18% of the conclusions concerning row hypotheses were incorrect and 9% concerning column hypotheses were incorrect when correlations were less than 10%. These respective figures were 22% and 2.5% for the 4 x 3 tables, 11% and 4% for the 5 x 2 tables, and 17% and 23% for the 5 x 4 tables. Most of these percentages would be considered outside the range of tolerance.

In Table 5, however, we see that if we do take into account the unequal variances but ignore small correlations then the percentage of errors in conclusions is reduced to within practical tolerance limits. The respective figures are now 6% and 5% for the 3 x 3 tables, 5% and 0.0% for the 4 x 3 tables, 2.5% and 0.0% for the 5 x 2 tables, and 2.5% and 2.5% for the 5 x 4 tables.

The significance of this result lies in the fact that variance and covariance estimation is quite tedious and time consuming in complex surveys. Therefore, any reduction in the number of calculations is usually worthwhile. Also, since the proposed method of analysis involves the inversion of the covariance matrix, if it can be reduced to diagonal form the accuracy of the analysis would be improved, especially for large numbers of observations.

6. Concluding Remarks

In concluding, I would like to say that this paper is written from a very practical point of view and any conclusions drawn should be interpreted with this in mind. Especially I would like to re-emphasize the limitations and restrictions placed on the proposed test for interactions. At the same time, however, the emperical results presented in Tables 1-5 are in general supported by theory developed by Walsh3/ in 1947 in which he considered a special case of the preceding problem in which $\rho_{ij} = \rho$ for all i and j, and $\sigma_i^2 = \sigma^2$ for all i. Under these assumptions he found simple correction factors for the χ^2- and F-statistics calculated under the assumption of independence, i.e.,

$$\chi^2_{\text{true}} = \frac{1}{1 - \rho} \chi^2_{\text{independent}}$$

$$F_{\text{true}} = \frac{1 - \rho'}{1 - \rho''} F_{\text{independent}}$$

And these correction factors do in general support the emperical results presented in Tables 1-5.

FOOTNOTES

1/ E. T. Browne, Introduction to the Theory of Matrix Algebra

2/ Henry Scheffe, 1959, The Analysis of Variance

3/ John E. Walsh, "Concerning the Effect of Intraclass Correlation on Certain Significance Tests," Annals of Mathematical Statistics, 1947

	C_1	C_2	C_3
R_1	23.35	1.13	1.18
R_2	6.07	20.24	16.66
R_3	1.81	11.49	3.08

$$\Sigma = \begin{bmatrix} 129.57 & 1.46 & 0.02 & 1.15 & -0.99 & -0.99 & 0.24 & -0.21 & -1.19 \\ 1.46 & 79.34 & 0.87 & -0.97 & 2.02 & -0.06 & -0.37 & 1.72 & -0.21 \\ 0.02 & 0.87 & 2.07 & -0.48 & -0.25 & 0.20 & -0.01 & -1.19 & 0.11 \\ 1.15 & -0.97 & -0.48 & 148.28 & 0.84 & 0.17 & 0.26 & -0.08 & -0.30 \\ -0.99 & 2.02 & -0.25 & 0.84 & 96.94 & 1.68 & -0.54 & 3.50 & -0.71 \\ -0.99 & -0.06 & 0.20 & 0.17 & 1.68 & 3.32 & -0.54 & -1.09 & 1.33 \\ 0.24 & -0.37 & -0.01 & 0.26 & -0.54 & -0.54 & 1.02 & 0.55 & 1.02 \\ -0.21 & 1.72 & -1.19 & -0.08 & 3.50 & -1.09 & 0.55 & 8.43 & 1.70 \\ -1.19 & -0.21 & 0.11 & -0.30 & -0.71 & 1.33 & 1.02 & 1.70 & 156.07 \end{bmatrix}$$

$$\Sigma_1 = \begin{bmatrix} 129.57 & 1.53 & 0.02 & 1.24 & -1.08 & -1.09 & 0.26 & -0.23 & -1.36 \\ 1.53 & 79.34 & 1.01 & -1.13 & 2.37 & -0.07 & -0.45 & 2.08 & -0.26 \\ 0.02 & 1.01 & 2.07 & -0.59 & -0.31 & 0.25 & -0.01 & -1.53 & 0.14 \\ 1.24 & -1.13 & -0.59 & 148.28 & 1.10 & 0.22 & 0.34 & -0.11 & -0.41 \\ -1.08 & 2.37 & -0.31 & 1.10 & 96.94 & 2.30 & -0.74 & 4.88 & -1.00 \\ -1.09 & -0.07 & 0.25 & 0.22 & 2.30 & 3.32 & -0.77 & -1.56 & 1.92 \\ 0.26 & -0.45 & -0.01 & 0.34 & -0.74 & -0.77 & 1.02 & 0.81 & 1.50 \\ -0.23 & 2.08 & -1.53 & -0.11 & 4.88 & -1.56 & 0.81 & 8.43 & 2.51 \\ -1.36 & -0.26 & 0.14 & -0.41 & -1.00 & 1.92 & 1.50 & 2.51 & 156.07 \end{bmatrix}$$

SOURCE	SUM OF SQUARES	DF	MEAN SQUARE	F-RATIO
ROW	48.20804	2	24.10402	16.10687
COLUMN	10.75986	2	5.37993	3.59500
ERROR	5.98602	4	1.49650	

$$\Sigma_2 = \begin{bmatrix} 129.57 & 0.44 & 0.18 & 0.34 & -0.22 & -0.25 & 0.16 & -0.37 & -0.23 \\ 0.44 & 79.34 & 0.19 & -0.28 & 0.41 & -0.13 & -0.07 & 0.52 & -0.16 \\ 0.18 & 0.19 & 2.07 & -0.36 & -0.22 & 0.34 & -0.14 & -0.23 & 0.23 \\ 0.34 & -0.28 & -0.36 & 148.28 & 0.75 & 0.31 & 0.19 & -0.55 & -0.32 \\ -0.22 & 0.41 & -0.22 & 0.75 & 96.94 & 0.81 & -0.27 & 0.61 & -0.41 \\ -0.25 & -0.13 & 0.34 & 0.31 & 0.81 & 3.32 & -0.16 & -0.32 & 0.48 \\ 0.16 & -0.07 & -0.14 & 0.19 & -0.27 & -0.16 & 1.02 & 0.21 & 0.17 \\ -0.37 & 0.52 & -0.23 & -0.55 & 0.61 & -0.32 & 0.21 & 8.43 & 0.63 \\ -0.23 & -0.16 & 0.23 & -0.32 & -0.41 & 0.48 & 0.17 & 0.63 & 156.07 \end{bmatrix}$$

SOURCE	SUM OF SQUARES	DF	MEAN SQUARE	F-RATIO
ROW	50.66118	2	25.33058	17.40663
COLUMN	9.23855	2	4.61928	3.17427
ERROR	5.82090	4	1.45522	

$$\Sigma_3 = \begin{bmatrix} 129.57 & 0.23 & 0.10 & 0.18 & -0.12 & -0.14 & 0.09 & -0.22 & -0.14 \\ 0.23 & 79.34 & 0.12 & -0.18 & 0.26 & -0.08 & -0.05 & 0.35 & -0.11 \\ 0.10 & 0.12 & 2.07 & -0.25 & -0.15 & 0.25 & -0.10 & -0.17 & 0.17 \\ 0.18 & -0.18 & -0.25 & 148.28 & 0.58 & 0.25 & 0.15 & -0.45 & -0.27 \\ -0.12 & 0.26 & -0.15 & 0.58 & 96.94 & 0.69 & -0.23 & 0.54 & -0.37 \\ -0.14 & -0.08 & 0.25 & 0.25 & 0.69 & 3.32 & -0.15 & -0.30 & 0.46 \\ 0.09 & -0.05 & -0.10 & 0.15 & -0.23 & -0.15 & 1.02 & 0.20 & 0.17 \\ -0.22 & 0.35 & -0.17 & -0.45 & 0.54 & -0.30 & 0.20 & 8.43 & 0.63 \\ -0.14 & -0.11 & 0.17 & -0.27 & -0.37 & 0.46 & 0.17 & 0.63 & 156.07 \end{bmatrix}$$

SOURCE	SUM OF SQUARES	DF	MEAN SQUARE	F-RATIO
ROW	48.69180	2	24.34590	16.81972
COLUMN	9.18756	2	4.59378	3.17368
ERROR	5.78985	4	1.44746	

$$\Sigma_4 = \begin{bmatrix} 129.57 & 0.46 & 0.19 & 0.36 & -0.24 & -0.28 & 0.17 & -0.41 & -0.26 \\ 0.46 & 79.34 & 0.22 & -0.33 & 0.49 & -0.16 & -0.09 & 0.63 & -0.20 \\ 0.19 & 0.22 & 2.07 & -0.45 & -0.27 & 0.43 & -0.18 & -0.30 & 0.30 \\ 0.36 & -0.33 & -0.45 & 148.28 & 0.99 & 0.41 & 0.25 & -0.74 & -0.44 \\ -0.24 & 0.49 & -0.27 & 0.99 & 96.94 & 1.11 & -0.37 & 0.86 & -0.58 \\ -0.28 & -0.16 & 0.43 & 0.41 & 1.11 & 3.32 & -0.23 & -0.47 & 0.70 \\ 0.17 & -0.09 & -0.18 & 0.25 & -0.37 & -0.23 & 1.02 & 0.30 & 0.26 \\ -0.41 & 0.63 & -0.30 & -0.74 & 0.86 & -0.47 & 0.30 & 8.43 & 0.93 \\ -0.26 & -0.20 & 0.30 & -0.44 & -0.58 & 0.70 & 0.26 & 0.93 & 156.07 \end{bmatrix}$$

SOURCE	SUM OF SQUARES	DF	MEAN SQUARE	F-RATIO
ROW	52.60782	2	26.30391	18.02400
COLUMN	9.42238	2	4.71119	3.22821
ERROR	5.83753	4	1.45938	

TABLE 1
(Number of Conclusions Changed When Original Data Had Non-Zero Correlations and Unequal Variances)

Size of Table Analyzed	Number of Data Sets	ERROR IN $\hat{\Sigma}$		INDEPENDENCE AND EQUAL VARIANCE	
		Row Hypothesis	Column Hypothesis	Row Hypothesis	Column Hypothesis
2 x 2	82	1	1	9	15
3 x 3	101	12	3	50	28
4 x 3	50	0	1	10	15
5 x 2	42	2	0	8	4
5 x 4	60	1	1	16	6

TABLE 2
(Average Absolute Percentage Change in F-Statistic for Data of Table 1)

Size of Table Analyzed	Number of Data Sets	ERROR IN $\hat{\Sigma}$		INDEPENDENCE AND EQUAL VARIANCE	
		Row Hypothesis	Column Hypothesis	Row Hypothesis	Column Hypothesis
2 x 2	82	3.0	4.0	72.0	91.5
3 x 3	101	13.0	9.0	91.0	83.0
4 x 3	50	3.0	3.0	73.0	80.0
5 x 2	42	4.0	8.5	46.5	92.0
5 x 4	60	4.0	7.0	45.0	48.0

TABLE 3
(Equal Variances Required and Small Correlations Assumed to be Zero)

Size of Table Analyzed	Number of Data Sets	Correlation Coefficient (ρ_{ij})	NUMBER OF CONCLUSIONS CHANGED	
			Row Hypothesis	Column Hypothesis
3 x 3	120	$\leq .05$	4	2
	120	$\leq .10$	5	2
	120	$\leq .25$	9	3
4 x 3	40	$\leq .05$	0	0
	40	$\leq .10$	1	0
	40	$\leq .25$	2	1
5 x 4	80	$\leq .05$	0	0
	80	$\leq .10$	0	0
	80	$\leq .25$	0	0

TABLE 4
(Unequal Variances Not Taken into Account When Small Correlations Assumed to be Zero)

Size of Table Analyzed	Number of Data Sets	Correlation Coefficient (ρ_{ij})	NUMBER OF CONCLUSIONS CHANGED	
			Row Hypothesis	Column Hypothesis
3 x 3	126	$\leq .05$	15	4
	124	$\leq .10$	22	11
	42	$\leq .25$	4	3
4 x 3	42	$\leq .05$	7	2
	42	$\leq .10$	9	1
	42	$\leq .25$	14	4
5 x 2	84	$\leq .05$	11	2
	84	$\leq .10$	9	3
	42	$\leq .25$	3	2
5 x 4	84	$\leq .05$	17	14
	105	$\leq .10$	18	24
	42	$\leq .25$	15	9

TABLE 5
(Unequal Variances Taken into Account When Small Correlations Assumed to be Zero)

Size of Table Analyzed	Number of Data Sets	Correlation Coefficient (ρ_{ij})	NUMBER OF CONCLUSIONS CHANGED	
			Row Hypothesis	Column Hypothesis
3 x 3	80	$\leq .05$	3	3
	80	$\leq .10$	5	4
	80	$\leq .25$	7	4
4 x 3	40	$\leq .05$	2	0
	40	$\leq .10$	2	0
	40	$\leq .25$	4	1
5 x 2	40	$\leq .05$	2	0
	40	$\leq .10$	1	0
	40	$\leq .25$	4	0
5 x 4	80	$\leq .05$	2	0
	80	$\leq .10$	2	2
	80	$\leq .25$	7	7

ESTIMATORS IN MULTIPLE FRAME SURVEYS

Richard E. Lund

Iowa State University and
Centro de Estadistica y Cálculo, Chapingo, México

Estimators appropriate for multiple frame surveys were proposed by Hartley [3]. This paper suggests an alteration in these estimators consisting of basing the weights associated with the sample from each frame upon the actual sample sizes obtained. The resulting estimators have equal or greater efficiency. Complexity is reduced in both the estimator and sample allocation determination.

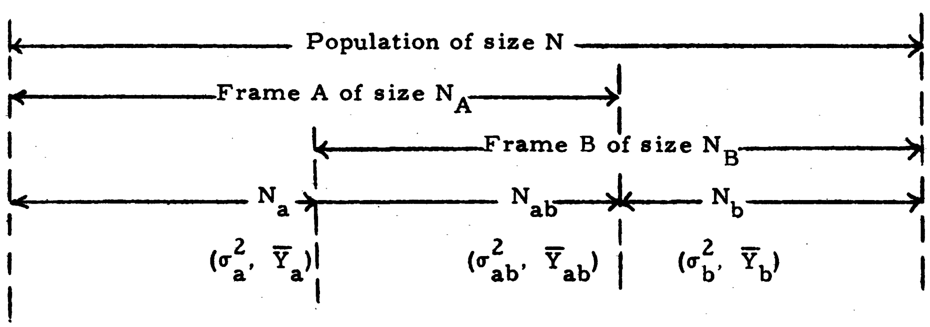
Introduction

A sampling frame or list is the keystone around which a sampling process is constructed. But often a single list corresponding to all elements in the desired population is not available. Consequently, two or more lists are frequently used to construct a frame of satisfactory coverage. In other situations a single suitable, but relatively costly, frame is available, however efficiency considerations suggest the joint use of another less costly incomplete frame.

Commonly used sampling estimators require the elimination of duplicated elements from the frame. However, such an elimination can be impractical. Hartley [3] proposed estimators and allocation formula suitable for use with two overlapping frames. This paper suggests alterations in these estimators which improve efficiency and decrease complexity.

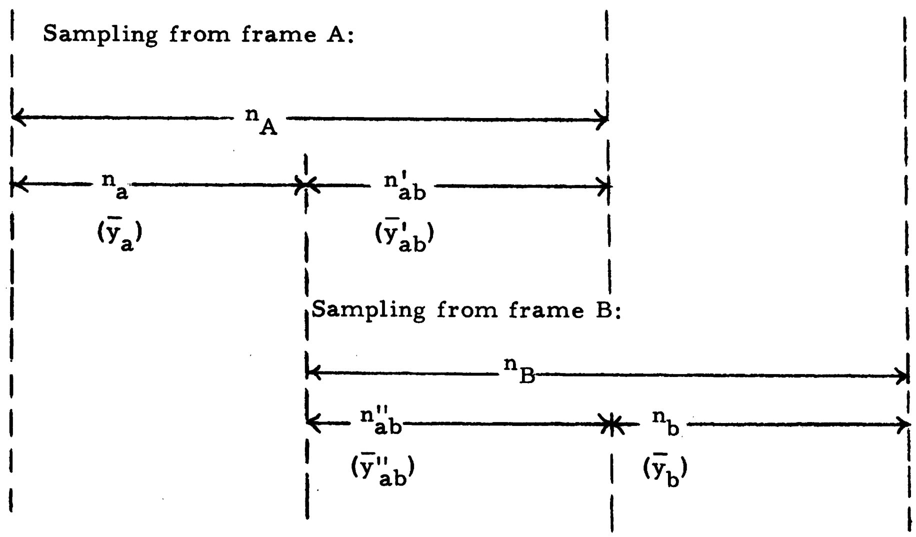
Notation and Nomenclature

The survey objective is considered to be the estimation of the total (ΣY_i) of characteristic "Y" for a population containing N elements. Complete coverage of the desired population is provided by two overlapping frames A and B of sizes N_A and N_B. The population can be separated into three domains: (a) the non-duplicated elements associated with frame A, (ab) the elements duplicated in both A and B, and (b) the non-duplicated elements in B. Using N_a, N_{ab} and N_b to represent the size of each domain, $\overline{Y}_a$, $\overline{Y}_{ab}$ and $\overline{Y}_b$, and σ_a^2, σ_{ab}^2 and σ_b^2 to represent the population means and unit variances, the basis can be presented schematically as:



Several expressions are simplified by referring to the relative size of the overlap with respect to frame size $(\alpha = N_{ab}/N_a$ and $\beta = N_{ab}/N_B)$.

Random samples of n_A and n_B elements are selected independently from the two frames. The division of the sample for any frame between the two domains (unduplicated and duplicated elements) is not considered to be subject to control by the sampler. Symbols n_a and n'_{ab} refer to the sample sizes resulting from the random division of n_A. Sizes n''_{ab} and n_b are associated similarly with the sample from frame B. Defining the sample means $\bar{y}_a$, $\bar{y}'_{ab}$, $\bar{y}''_{ab}$ and $\bar{y}_b$, we have:



Having this basis, attention can be turned to two principal cases: first, the case of domain sizes N_a, N_{ab} and N_b known and second, the case of N_a, N_{ab} and N_b unknown but N_A and N_B known with estimates of α and β available for sample allocation.

Case of N_a, N_{ab} and N_b Known

Hartley suggested the unbiased estimator of the population total

$$\hat{Y}_H = N_a\bar{y}_a + N_{ab}p\bar{y}'_{ab} + N_{ab}(1-p)\bar{y}''_{ab} + N_b\bar{y}_b \quad (1)$$

where $o \leq p \leq 1$. The variance of this estimator is approximately

$$\begin{aligned} \operatorname{Var}(\hat{Y}_H) &\doteq \frac{N_A^2}{n_A}\left[(1-\alpha)\sigma_a^2 + \alpha p^2\sigma_{ab}^2\right] \\ &+ \frac{N_B^2}{n_B}\left[(1-\beta)\sigma_b^2 + \beta(1-p)^2\sigma_{ab}^2\right]. \quad (2) \end{aligned}$$

Finite population corrections have been ignored.

Sampling costs can be expressed by the linear function

$$\text{Total Cost} = n_A c_A + n_B c_B \quad (3)$$

where c_A and c_B define unit costs of sampling from each frame respectively. The problem of optimizing the sample allocation among the two frames and finding the optimum value for p consists of minimizing (2) as a function of n_A, n_B and p subject to restriction (3). Hartley expressed the value for p as a bi-quadratic while additional formulas were given for n_A and n_B. The solution for p, however, has the simple expression

$$p_o = \frac{\alpha n_A}{\alpha n_A + \beta n_B} \quad (4)$$

Thus, the optimum value for p is the ratio of the expected value of n'_{ab} with respect to the expected value of $n'_{ab} + n''_{ab}$.

Hartley's procedure does not consider the actual division achieved (at random) of the n_A and n_B elements among the domains. Thus, one may ask whether a gain is achieved by making p a function of n'_{ab} and n''_{ab}.

To reach a solution, the variance of (1) can be taken in two steps by use of the well known theorem expressed in symbols of the current problem

$$\begin{aligned} \operatorname{Var}(\hat{Y}) &= E[\operatorname{Var}(\hat{Y} \mid n'_{ab}, n''_{ab})] \\ &+ \operatorname{Var}[E(\hat{Y} \mid n'_{ab}, n''_{ab})] \quad (5) \end{aligned}$$

where the condition, $|n'_{ab}, n''_{ab})$, represents the actual sample counts achieved. The variable within the second term, $E[(\hat{Y} \mid n'_{ab}, n''_{ab})]$, equals the population total for any value of p. Thus, its variance equals zero.

Disregarding finite population corrections,

$$\begin{aligned} \operatorname{Var}(\hat{Y} \mid n'_{ab}, n''_{ab}) &= \frac{N_a^2}{n_a}\sigma_a^2 + p^2\frac{N_{ab}^2}{n'_{ab}}\sigma_{ab}^2 \\ &+ (1-p)^2\frac{N_{ab}^2}{n''_{ab}}\sigma_{ab}^2 + \frac{N_b^2}{n_b}\sigma_b^2. \quad (6) \end{aligned}$$

Minimization of (6) as a function of p gives the solution

$$p_o = \frac{n'_{ab}}{n'_{ab} + n''_{ab}} \quad . \quad (7)$$

The variance for the estimator with this value for p can be found by substituting (7) into (6) and obtaining the expected value. The estimator and its approximate variance are

$$\hat{Y}_L = N_a\bar{y}_a + N_{ab}\bar{y}_{ab} + N_b\bar{y}_b \qquad (8)$$

where $$\bar{y}_{ab} = \frac{n'_{ab}\bar{y}'_{ab} + n''_{ab}\bar{y}''_{ab}}{n'_{ab} + n''_{ab}}$$

and

$$\mathrm{Var}(\hat{Y}_L) \doteq \frac{N_A^2}{n_A}(1-\alpha)\sigma_a^2 + \frac{N_A N_B \alpha\beta}{\alpha n_A + \beta n_B}\sigma_{ab}^2 + \frac{N_B^2}{n_B}(1-\beta)\sigma_b^2 \ . \qquad (9)$$

The order of the approximation in (9) is the same as for (2).

While it can be proven that estimator (8) is always equal or greater in efficiency than (1), this increase in efficiency is not reflected in variance approximation (9). Substitution of (4) into (2) provides an expression identical to (9) which indicates that both estimators are equal in efficiency to the order of the approximation.

The departure of variance approximation (9) from the true variance of $\hat{Y}_L$ was estimated by examining terms through the second order of a Taylor's expansion of the random variables $1/n_a$, $1/(n'_{ab} + n''_{ab})$ and $1/n_b$ around the values $1/\alpha n_A$, $1/(\alpha n_A + \beta n_B)$ and $1/\beta n_B$ respectively. The first term of (9) could be corrected by multiplying by $[1 + \alpha/(1-\alpha)n_A]$ and the third term by $[1 + \beta/(1-\beta)n_B]$. The correction for the second term is $[1 + \delta/(\alpha n_A + \beta n_B)]$ where δ is a weighted average of $(1-\alpha)$ and $(1-\beta)$, the weights being αn_A and βn_B. As noted earlier, approximation (9) is also appropriate for the variance of estimator (1) with p expressed by (4), but δ in the second term correction becomes the sum of $(1-\alpha)$ and $(1-\beta)$. No change occurs in the correction for the first and third terms. Thus, it is seen that variance approximation (9) (or 2) is reasonably accurate for all but very small samples and in addition, the gain in efficiency by use expression (7) for p instead of (4) is negligible except for extremely small samples.

The general solution of the allocation problem, as found by minimizing (9) as a function of n_A and n_B subject to cost equation (3) can be expressed by the iterative system

$$r_1 = \sqrt{\frac{c_B}{c_A}\left(\frac{\beta}{\alpha}\right)}$$

$$r_{i+1}^2 = \frac{c_B}{c_A}\left(\frac{\beta}{\alpha}\right)^2 \times \frac{(r_i + \frac{\beta}{\alpha})^2(1-\alpha)\sigma_a^2 + r_i^2\sigma_{ab}^2}{(r_i + \frac{\beta}{\alpha})^2(1-\beta)\sigma_b^2 + (\frac{\beta}{\alpha})^2\beta\sigma_{ab}^2} \qquad (10)$$

where $r = n_A/n_B$. Practice with the system for several values of the parameters indicated that few iterations are required in most cases.

To determine the sensitivity of the estimator to deviations from optimum allocation, the optimum value for r and the corresponding variance were computed for a wide range of values of the parameters. A comparison was made to the variances corresponding to deviations of ten percent in both directions from the optimum (that is, $0.90r_o$ and $1.10r_o$). The variance was increased by more than one percent in very few instances by the deviation from optimum allocation.

The case of complete coverage by a relatively costly frame merits some additional consideration. Defining A as the complete frame, it is clear that $N_b = 0$ and $\beta = 1$ which enable a simple graphical presentation of the optimum allocation. Figure 1 displays the solutions for four relative cost levels and three variance ratios.

Case of N_a, N_{ab} and N_b Unknown

Estimators (8) can serve as the starting point for the case of unknown N_a, N_{ab} and N_b. However, it is necessary to insert estimates for these sizes by use of sample data and known N_A and N_B. The expressions $N_A(n_a/n_A)$ and $N_B(n_b/n_B)$ are unbiased estimators of N_a and N_b, respectively. Two unbiased estimators of N_{ab} are available: $N_A(n'_{ab}/n_A)$ and $N_B(n''_{ab}/n_B)$. Using p and (1-p) as undetermined weights for the two estimates of N_{ab} and substituting these expressions in (8), an unbiased estimator for the population total is

$$\dot{Y} = \frac{N_A}{n_A}n_a\bar{y}_a + \left[\frac{N_A}{n_A}n'_{ab}p + \frac{N_B}{n_B}n''_{ab}(1-p)\right]\bar{y}_{ab} + \frac{N_B}{n_B}n_b\bar{y}_b \qquad (11)$$

where as before $\bar{y}_{ab}$ is the sample mean of all elements selected from the domain of duplicated elements.

Use of theorem (5) provides the approximate variance

$$\mathrm{Var}(\dot{Y}) \doteq \frac{N_A^2}{n_A}(1-\alpha)\sigma_a^2 + \frac{N_A N_B \alpha\beta}{\alpha n_A + \beta n_B}\sigma_{ab}^2 + \frac{N_B^2}{n_B}(1-\beta)\sigma_b^2 + \frac{N_A^2(1-\alpha)\alpha}{n_A}[\bar{Y}_a - p\bar{Y}_{ab}]^2 + \frac{N_B^2(1-\beta)\beta}{n_B}[\bar{Y}_b-(1-p)\bar{Y}_{ab}]^2 . \quad (12)$$

The final two terms represent the increase in variance due to not knowing N_a, N_{ab} and N_b. These terms make a significant contribution unless either the overlap is nearly complete or is relatively small.

The degree of the approximation in variance (12) is not the same as noted earlier for (9) or (2). All terms are exact except for the second. Examination of terms through the second order of a Taylor's expansion of the second term in $\mathrm{Var}(\dot{Y}|n'_{ab}, n''_{ab})$ suggests a multiplicative correction equal to or less than $[1+(1-\alpha)/\alpha n_A + (1-\beta)/\beta n_B]$. Thus, the approximation is reasonably accurate for all practical uses.

Minimization of (12) as a function of p, n_A and n_B subject to cost equation (3) specifies

$$p_o = \frac{\frac{N_A(1-\alpha)}{n_A}\bar{Y}_a + \frac{N_B(1-\beta)}{n_B}(\bar{Y}_{ab}-\bar{Y}_b)}{\left[\frac{N_A(1-\alpha)}{n_A} + \frac{N_B(1-\beta)}{n_B}\right]\bar{Y}_{ab}} \quad (13)$$

and provides the sample allocation among the two frames. The sample allocation is expressed again as an iterative system

$$r_1 = \sqrt{\frac{c_B}{c_A}\left(\frac{\beta}{\alpha}\right)}$$

$$r_{i+1}^2 = \frac{c_B}{c_A}\left(\frac{\beta}{\alpha}\right)^2\left[(1-\alpha)\sigma_a^2 + \frac{r_i^2\alpha\sigma_{ab}^2}{(r_i+\frac{\beta}{\alpha})^2} + \frac{r_i^2\alpha(1-\alpha)(\bar{Y}_a+\bar{Y}_b-\bar{Y}_{ab})^2}{[r_i+\frac{\beta}{\alpha}(\frac{1-\alpha}{1-\beta})]^2}\right]$$

$$\times\left[(1-\beta)\sigma_b^2 + \frac{(\frac{\beta}{\alpha})^2\beta\sigma_{ab}^2}{(r_i+\frac{\beta}{\alpha})^2} + \frac{[\frac{\beta}{\alpha}(\frac{1-\alpha}{1-\beta})]^2\beta(1-\beta)(\bar{Y}_a+\bar{Y}_b-\bar{Y}_{ab})^2}{[r_i+\frac{\beta}{\alpha}(\frac{1-\alpha}{1-\beta})]^2}\right]^{-1} . \quad (14)$$

Use of the system with various values of the parameters indicates that normally few iterations are required.

It is apparent that the sample allocation is based on preliminary estimates of the unit variances and population means as well as upon preliminary estimates of N_a, N_{ab} and N_b as expressed in α and β. A numerical investigation similar to that discussed for the case of N_a, N_{ab} and N_b known, indicates the efficiency of the estimator is rather insensitive to moderate departures from optimum allocation.

To determine the value of p, one may, of course, utilize the estimates of the parameters used in determining the sample allocation. However, some gain in efficiency would probably be achieved by using the sample data. An estimator of optimum p from the sample data is

$$\hat{p} = \frac{\frac{N_A n_a}{n_A^2}\bar{y}_a + \frac{N_B n_b}{n_B^2}(\bar{y}_{ab}-\bar{y}_b)}{\left(\frac{N_A n_a}{n_A^2} + \frac{N_B n_b}{n_B^2}\right)\bar{y}_{ab}} . \quad (15)$$

Of course, use of (15) disturbs the unbiasedness of estimator (11) for p is now a function of n'_{ab} and n''_{ab}. However, the degree of bias is considered to be negligible. The expected value of estimator (11) with p given by (15) was approximated by use of a Taylor's expansion in which terms through the second order were retained. This approximation suggested the second term will be biased downward by the multiplicative factor $[1-\delta]$ where δ is the weighted mean of $(1/n_A)$ and $(1/n_B)$, the weights being $N_A(1-\alpha)/n_A$ and $N_B(1-\beta)/n_B$.

Hartley suggested an estimator to handle the current case, but his weight variable p was used to weight two individual estimates of the total over the duplicated elements ($N_{ab}\bar{Y}_{ab}$). The expression for the optimum value for such a p includes the parameters σ_a^2, σ_{ab}^2 and σ_b^2 . Thus, the estimator suggested herein is somewhat simplier. But in addition it can be proven that estimator (11) always has equal or greater efficiency than the estimator suggested by Hartley.

The reduction in variance can be quite important. In the not unusual case of equal per unit variances, equal means among the various strata, coefficient of variation of .10, $\alpha = \beta = .80$, and relative costs of $c_A/c_B = 4$, the reduction in variance by use of (11) was about one-fourth. Cases requiring more extreme sample allocation between the two frames, by reason for example of greater cost differences, show a greater gain in efficiency by use of (11).

Figure 2 presents the optimum allocation for the special case of $\alpha = \beta$, equal unit variances and approximately equal domain means. Four relative cost levels and three values for the coefficient of variation are considered.

Numerical Example

It may be assumed that a survey is to be conducted to measure characteristics of daily milk consumption in Mexico City. The sampling unit selected is the household. Among the various sampling frames available, it is decided to use the following two:

1. Telephone directory
2. Housing registration records maintained by the Secretary of Housing.

Preliminary investigations indicate the second frame covers the desired population satisfactorily. While the first frame provides inadequate coverage, costs of using it are substantially less because the data can be collected by telephone. It is assumed to be more economical to collect all data by personal interview for households selected from the second frame by reason of difficulties in matching housing unit records to telephone numbers.

It is estimated that the telephone directory provides 40 percent coverage. Per unit variances are assumed to be equal in both the duplicated parts of the population. Applying this information to Figure 1 suggests the sample allocation of 50 percent to each frame. If the relative cost ratio of the housing records with respect to the telephone directory were only 2, the optimum allocation would be to assign only 20 percent to the lower cost frame.

As an example of different conditions, suppose that a similar survey is planned for a U.S. city. The investigator decides to use two frames similar to those suggested earlier. However, in this case preliminary investigations show that each frame covers about 90 percent of the desired population while complete coverage is provided by using both frames. The 90 percent value is only a crude estimate for survey planning purposes and the investigator wishes to treat the domain sizes as unknown (N_a, N_{ab} and N_b). Supposing that the per unit costs of using the telephone directory are only one-half those of using a housing unit list and the variances and means for the various domains are equal approximately with a coefficient of variation of 0.5, the optimum allocation indicated by Figure 2 is to assign 75 percent of the sample to the lower cost frame.

It is clear that many practical situations cannot be handled by Figures 1 and 2. The optimum allocation for these can be determined by iterative systems (10) or (14).

References

[1] Cochran, Robert S. (1964). "Multiple Frame Sample Surveys." Proceedings of Social Science Section of American Statistical Association meetings, Chicago, Illinois.

[2] Cochran, Robert S. (1967). "The Estimation of Domain Sizes When Sampling Frames Are Interlocking." Mimeographed paper at the American Statistical Association meetings, Social Science Section, Washington, D.C.

[3] Hartley, H.O. (1962). "Multiple Frame Surveys." Proceedings of the Social Science Section of American Statistical Association meetings, Minneapolis, Minnesota.

[4] Steinberg, J. (1965). "A Multiple Frame Survey for Rare Population Elements." Proceedings of the Social Science Section of American Statistical Association meetings.

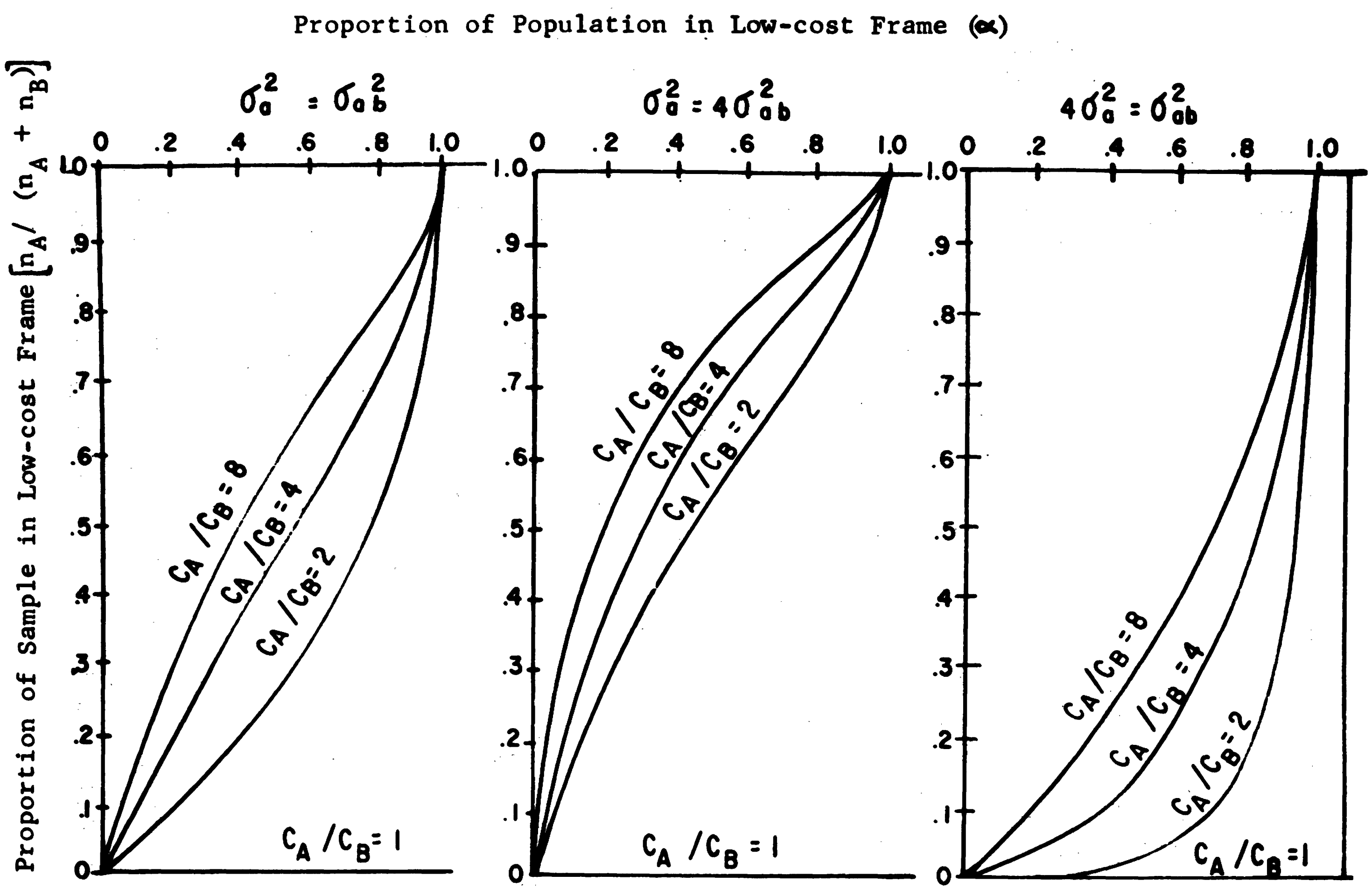


Figure 1

Optimum Allocation for Special Case of Complete Coverage by Costly Frame

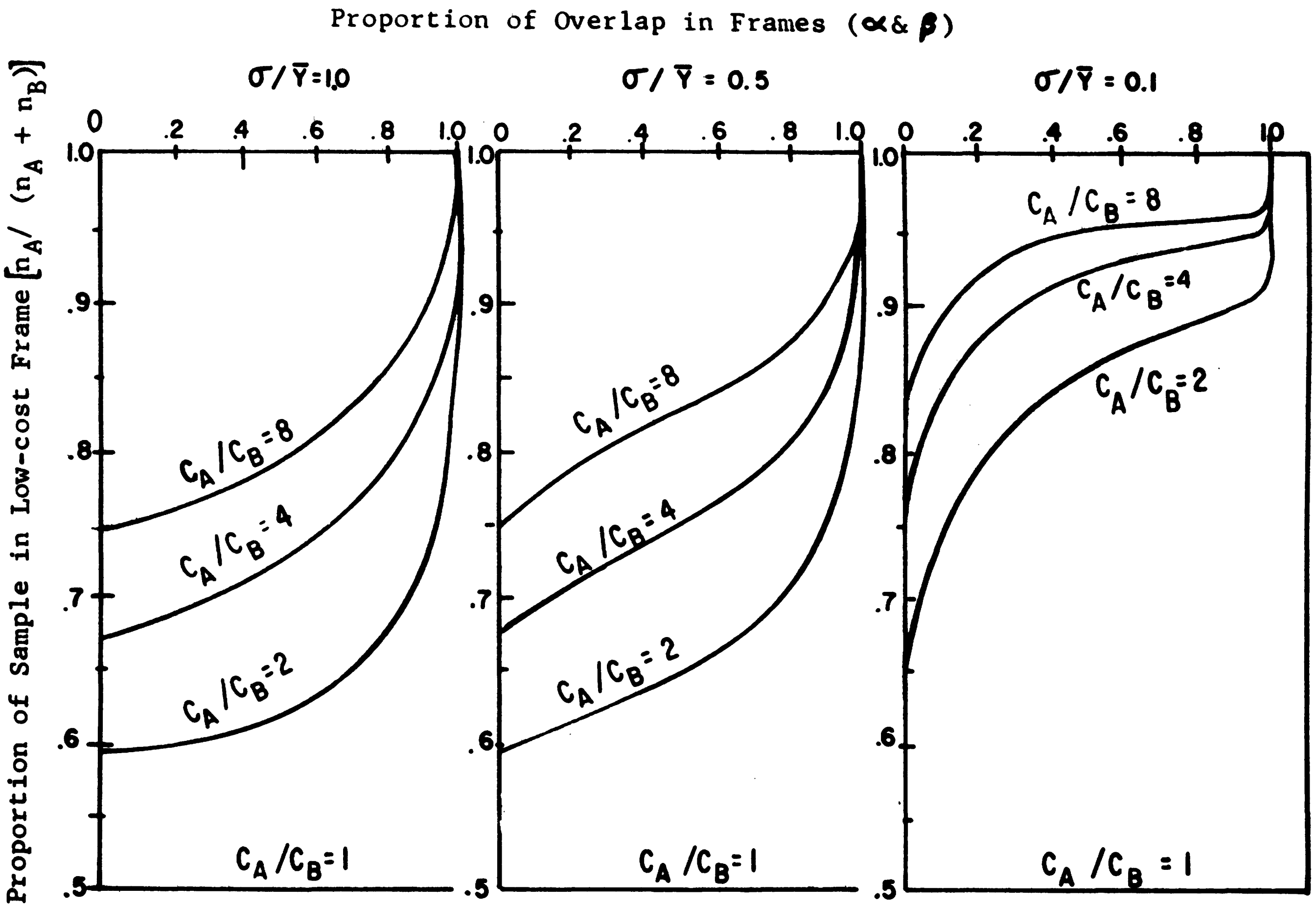


Figure 2

Optimum Allocation for Special Case of Equal Frame Sizes, Equal Variances and Means (Approx.), and Unknown Domain Sizes (N_a, N_{ab}, N_b)

MULTIPLE CURVILINEAR CORRELATION AND REGRESSION ANALYSIS THROUGH POLYNOMIAL DATA TRANSFORMATION

Roy W. Meadows, University of Evansville

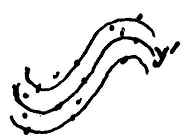
While data transformations are commonplace in multiple correlation analysis, and least squares curve fitting is routine throughout science, to the best of my knowledge, the use of the latter to achieve the former is not in widespread use. A computer is of course required to make such a procedure feasible for routine use, and a computer program by the author of this paper was accepted for distribution by the IBM library of contributed programs in September of 1967.[1] There was no other such program in the IBM 1620 library at that time, and I find no description of such a procedure in current elementary texts. A search of the articles in the Journal of the American Statistical Association dating from 1956 produced no titles suggestive of the procedure to be described here.

In the computer program cited, a least squares cubic polynomial was used. An explanation for the procedure is stated in the form of a simple mathematical theorem, using the case of the general polynomial. The assumptions are as important to note as the conclusion.

Theorem: If a set of data is generated by $y = a_0 + a_1x + a_2x^2 + \dots + a_nx^n + \varepsilon$, then $y' = a_0 + a_1x + a_2x^2 + \dots + a_nx^n$, the best fitting curve to the data, is linearly related to y.

Proof. Substitute y' for the polynomial in the equation for y and obtain $y = y' + \varepsilon$. Therefore y and y' are identical except for ε and are thus linearly related.

Geometrically, y and y' would be parallel polynomials through each ε as shown and thus linear with respect to each other. Another interpretation would be that y' maps the data onto the y-axis where it is linearly related to the y values. A similar straightening of data occurs when suitable data are plotted on log, semi-log, or probability paper. The use of polynomials to approximate other functions is as common as series computation. The procedure used here might be thought of as data straightening through plotting data on universal polynomial paper, a process facilitated by the computer.



To the extent that the hypothesis of the theorem is satisfied, i.e., to the extent that the polynomial in y fits the function generating the data, the transformation y' will produce linear data. In the examples to be discussed below, the fit and the resulting linearity range from fair to perfect.

In the computer program mentioned above the cubic least squares polynomial was chosen rather than an even powered polynomial because the cubic equation tends in a linear way beyond the sample data points, making some extrapolation possible. The cubic rather than the quintic was chosen because the computer I was using does not have enough core storage to solve the normal equations for 13 variables in Fortran. There is also the possibility that roundoff error could become quite a nuisance with the quintic equation. The usual normal equations for fitting $y' = a_0 + a_1x + a_2x^2 + a_3x^3$ to the data are:

$$\begin{aligned}
a_0N + a_1\Sigma x + a_2\Sigma x^2 + a_3\Sigma x^3 &= \Sigma y \\
a_0\Sigma x + a_1\Sigma x^2 + a_2\Sigma x^3 + a_3\Sigma x^4 &= \Sigma xy \\
a_0\Sigma x^2 + a_1\Sigma x^3 + a_2\Sigma x^4 + a_3\Sigma x^5 &= \Sigma x^2y \\
a_0\Sigma x^3 + a_1\Sigma x^4 + a_2\Sigma x^5 + a_3\Sigma x^6 &= \Sigma x^3y
\end{aligned}$$

The simple correlation matrices for a set of specimen data correlated before the transformations and correlated after the transformations, together with the cubic polynomial constants, a_0 to a_3, for the relationship between the dependent and each independent variable are shown in Table 1. The functions which were used to generate the data are shown in corresponding positions between the correlation matrices.

In Table 2, the raw and the transformed data are given as printed from the cards, with the dependent variable listed last on each second line (the zeros on each second line of the transformed data were produced by unused positions in the transformation program). As listed between the correlation matrices, the functions producing the data are Arctan x, a parabola, a semicircle, $\sqrt{x}$, etc. Plotting of the data will show the very considerable increase in linearity in most of it.

Except for the linear relationship, these data were prepared from functions often used for transformations, and in most cases the polynomial simulated the function rather well as shown by the near perfect correlations. The procedure performed the useful task of supplying the

1. "Transformations to Check and Increase Linearity in Data,"IBM 1620-.06.0.250,(Sept. 1967) by Roy W. Meadows. The program transforms up to 12 independent variables at a time.

Table 1

UNTRANSFORMED DATA 9

TOTAL NUMBER = 12

MEANS X(1) TO X(9)

4.60	1.11	5.24	10.55	3.20	1.88	.93	4.60	1.49

STANDARD DEVIATIONS X(1) TO X(9)

3.294	.469	3.664	8.221	3.288	.976	1.446	3.294	2.883

CORRELATION RATIOS X(1) TO X(9)

1.00000	.82239	-.04316	-.03981	.96202	.94522	.87370	1.00000	-.61963
.82239	1.00000	-.02889	-.06807	.65704	.96030	.99194	.82239	-.90332
-.04316	-.02889	1.00000	.44584	-.02753	-.04792	.01627	-.04316	-.20687
-.03981	-.06807	.44584	1.00000	-.01759	-.06136	-.08102	-.03981	.16529
.96202	.65704	-.02753	-.01759	1.00000	.82616	.73061	.96202	-.45700
.94522	.96030	-.04792	-.06136	.82616	1.00000	.97836	.94522	-.79604
.87370	.99194	.01627	-.08102	.73061	.97836	1.00000	.87370	-.90131
1.00000	.82239	-.04316	-.03981	.96202	.94522	.87370	1.00000	-.61963
-.61963	-.90332	-.20687	.16529	-.45700	-.79604	-.90131	-.61963	1.00000

y = Arctan x, $-\frac{2}{5}(x-5)^2+10$, $\sqrt{100-(x-10)^2}$, $\sqrt{x}$, sinh x , e^x , x , 1/x

CUBIC LEAST SQUARES POLYNOMIAL TRANSFORMED 9

TOTAL NUMBER = 12

MEANS X(1) TO X(9)

4.60	4.60	4.60	4.60	4.60	4.60	4.60	4.60	4.60

STANDARD DEVIATIONS X(1) TO X(9)

3.294	3.233	3.294	3.169	3.280	3.293	3.290	3.294	2.981

CORRELATION RATIOS X(1) TO X(9)

1.00000	.98170	.99999	.96214	.99596	.99993	.99896	1.00000	.90530
.98170	1.00000	.98178	.91254	.98294	.98283	.98870	.98170	.92902
.99999	.98178	1.00000	.96163	.99577	.99992	.99897	.99999	.90601
.96214	.91254	.96163	1.00000	.96604	.96143	.95603	.96214	.77656
.99596	.98294	.99577	.96604	1.00000	.99617	.99663	.99596	.89369
.99993	.98283	.99992	.96143	.99617	1.00000	.99913	.99993	.90409
.99896	.98870	.99897	.95603	.99663	.99913	1.00000	.99896	.91335
1.00000	.98170	.99999	.96214	.99596	.99993	.99896	1.00000	.90530
.90530	.92902	.90601	.77656	.89369	.90409	.91335	.90530	1.00000

CUBIC LEAST SQUARES TRANSFORMATION CONSTANTS

```
TR VR NO  1  COEF=-269.9937E-02 265.9642E-01-455.8080E-01 223.3441E-01   004
TR VR NO  2  COEF= 380.2200E-04 396.8686E-02-393.8266E-03-347.0773E-06   005
TR VR NO  3  COEF= 161.4088E-02 191.8828E-02-965.2864E-04 604.1991E-08   006
TR VR NO  4  COEF= 572.9356E-03 250.3119E-02-312.4604E-03 157.6915E-04   007
TR VR NO  5  COEF=-952.5550E-04 173.6754E-02-917.4105E-03 484.9071E-03   008
TR VR NO  6  COEF= 828.5387E-03 800.3786E-03 762.4955E-03 251.0349E-03   009
TR VR NO  7  COEF=    -.0000E-99 100.0000E-02    -.0000E-99    -.0000E-99   010
TR VR NO  8  COEF= 867.1686E-02-124.3719E-01 314.5788E-02-198.7872E-03   011
```

right function for a good transformation with a minimum of effort. Admittedly, the transformation for the reciprocal data falls about 19 per cent in variance short of the job of achieving perfect linearity and it is suggested that a higher degree polynomial might well do a better job. The linear case was included to show that the transformation does no harm to data that are already linear (except in the case of dichotomized data, to be discussed below) and that it actually serves as a verification of the linearity of the data in the case when the correlation coefficient shows little or no increase. Without knowing whether the data are linear or not, one cannot take the null hypothesis of no significant relationship and one cannot be sure that there is not a stronger curvilinear relationship in the data which is being wasted.

It should also be noted that all the

UNTRANSFORMED DATA Table 2 9

0.7900	9.7434	19.9499	.1000	.8850	.0000	1.0000
1.0000	1.0000					
1.1100	.5279	.2020	.4000	1.4480	.6932	2.0000
.5000	2.0000					
1.2500	9.1833	19.5394	.9000	1.8185	1.0986	3.0000
.3333	3.0000					
1.3300	1.1270	.8348	1.6000	2.0950	1.3863	4.0000
.2500	4.0000					
1.3750	8.5355	18.6603	2.5000	2.3125	1.6094	5.0000
.2000	5.0000					
1.4070	1.8378	2.0000	3.6000	2.4916	1.7918	6.0000
.1666	6.0000					
1.4295	7.7386	17.1414	4.9000	2.6433	1.9459	7.0000
.1429	7.0000					
1.4467	2.7639	4.0000	6.4000	2.7765	2.0794	8.0000
.1250	8.0000					
1.4600	6.5811	14.3589	8.1000	2.8933	2.1972	9.0000
.1111	9.0000					
1.4700	5.0000	10.0000	10.0000	2.9995	2.3026	10.0000
.1000	10.0000					
.1980	9.9497	.0004	.0040	.1990	-1.6090	.2000
5.0000	.2000					
.1000	.0100	19.9900	0.0010	.1000	-2.3030	.1000
10.0000	.1000					

1

CUBIC LEAST SQUARES POLYNOMIAL TRANSFORMED 9

.8759	.9980	1.5242	.8201	1.0593	.8285	1.0000	012
-.8185	1.0000	.0000	.0000	.0000	.0000		013
1.2072	2.0232	1.9977	1.5251	1.9682	1.8333	2.0000	014
3.2146	2.0000	.0000	.0000	.0000	.0000		015
2.9474	3.0022	2.2984	2.5841	2.9452	2.9609	3.0000	016
4.8684	3.0000	.0000	.0000	.0000	.0000		017
4.5901	4.0100	3.1486	3.8426	3.9754	4.0723	4.0000	018
5.7558	4.0000	.0000	.0000	.0000	.0000		019
5.7546	5.0047	3.8473	5.1242	5.0115	5.1381	5.0000	020
6.3084	5.0000	.0000	.0000	.0000	.0000		021
6.6967	5.9993	5.0656	6.2704	6.0372	6.1548	6.0000	022
6.6860	6.0000	.0000	.0000	.0000	.0000		023
7.4186	7.0045	6.1731	7.1912	7.0411	7.1228	7.0000	024
6.9580	7.0000	.0000	.0000	.0000	.0000		025
8.0046	7.9912	7.7453	7.9283	8.0334	8.0468	8.0000	026
7.1658	8.0000	.0000	.0000	.0000	.0000		027
8.4785	9.0004	9.2821	8.7280	8.9945	8.9310	9.0000	028
7.3284	9.0000	.0000	.0000	.0000	.0000		029
8.8470	9.9924	11.1555	10.1272	9.9461	9.7789	10.0000	030
7.4592	10.0000	.0000	.0000	.0000	.0000		031
.9525	.1959	1.6148	.5829	.2178	.4690	.2000	032
.2820	.2000	.0000	.0000	.0000	.0000		033
-.4737	.0776	1.4468	.5754	.0697	-.0369	.1000	034
.0913	.1000	.0000	.0000	.0000	.0000		035

simple r's with the dependent variable become positive under the transformations and that the positive-negative property of simple linear r's is lost. The simple r's with the dependent variable have become curvilinear r's and, like the multiple R, are positive. There are usually some very small negative r's among the intercorrelations for reasons to be discussed below.

An Example from Real Data.

Since I do not process a large volume of statistical data, I have not run across any really spectacular examples with real data of improved output due to transformations. We have found the transformation procedure to be useful, however, both in improved correlations and regression equations and in checking linearity. I am sure there would be some spectacular results occasionally where statistical data are processed in large volume.

The following example was among the best of 6 or 8 sets of data run during the year. The simple correlation matrices are shown in table 3, followed by the multiple R squared and plotbacks for both untransformed and transformed data in

Table 3

2 TO 14 VARIABLE CORRELATION MATRIX

DATA SET 2, UNTRANSFORMED 2

TOTAL NUMBER = 34

MEANS X(1) TO X(9)

490.82	3.12	2.22	2.22	2.57	3.80	5.31	5.98	4.70

STANDARD DEVIATIONS X(1) TO X(9)

85.654	.415	.709	.759	.687	2.243	2.100	2.316	2.541

CORRELATION RATIOS X(1) TO X(9)

.99999	.56379	.65776	.37149	.51373	.61732	.47854	.61046	.13142
.56379	.99999	.57397	.32621	.37250	.44498	.44942	.48468	.28451
.65776	.57397	1.00000	.57890	.75113	.45020	.29445	.52626	.04614
.37149	.32621	.57890	1.00000	.50433	.36180	.32114	.41795	-.02853
.51373	.37250	.75113	.50433	1.00000	.49547	.40271	.39584	.04701
.61732	.44498	.45020	.36180	.49547	.99999	.50623	.29087	.19425
.47854	.44942	.29445	.32114	.40271	.50623	1.00000	.54475	.17484
.61046	.48468	.52626	.41795	.39584	.29087	.54475	1.00000	.15267
.13142	.28451	.04614	-.02853	.04701	.19425	.17484	.15267	1.00000
.40339	.53095	.32465	.09138	.16682	.53154	.16812	.16145	.29818
.32751	.57520	.34596	.14175	.14527	.52360	.26854	.22931	.44937
.40816	.52531	.41946	.18666	.21527	.61559	.24283	.17619	.37702

DATA SET 2, TRANSFORMED 2

TOTAL NUMBER = 34

MEANS X(1) TO X(9)

490.82	490.82	490.82	490.82	490.82	490.82	490.82	490.82	490.82

STANDARD DEVIATIONS X(1) TO X(9)

85.654	49.286	58.417	32.527	47.404	53.346	41.184	54.821	26.643

CORRELATION RATIOS X(1) TO X(9)

.99999	.58013	.68200	.37968	.55342	.62278	.48080	.64000	.31099
.58013	1.00000	.67880	.33348	.38924	.43387	.45114	.51203	.38762
.68200	.67880	.99999	.54919	.73383	.47063	.38148	.58114	.19088
.37968	.33348	.54919	.99999	.53071	.38234	.35848	.44116	.20236
.55342	.38924	.73383	.53071	1.00000	.51086	.46093	.40265	-.00405
.62278	.43387	.47063	.38234	.51086	.99999	.51859	.35837	.23258
.48080	.45114	.38148	.35848	.46093	.51859	.99999	.58240	.24728
.64000	.51203	.58114	.44116	.40265	.35837	.58240	.99999	.20045
.31099	.38762	.19088	.20236	-.00405	.23258	.24728	.20045	.99998
.48075	.44141	.27214	-.00504	.08766	.56899	.35363	.21551	.30669
.52919	.50027	.44100	.23305	.40633	.59387	.41280	.27257	-.02732
.44448	.45037	.43205	.25354	.32262	.67703	.31319	.13733	.12141

Table 4 and Table 5.[2] In the simple correlation matrices, all the simple correlations between the dependent and the independent variables increased (the dependent variable is X(1) as before). It should be noticed that some of the intercorrelations decrease due to the transformations. This can be explained simply by observing that where there is considerable scatter in the data, linearity with respect to the dependent and an independent variable may not be linearity between two independent variables. In the specimen data presented first, it can be seen that this drop in intercorrelations is much less frequent because the

2. The program used was IBM 1620-.06.0.043, "Multiple Regression Package for the Card 1620," by Otto Dykstra, Jr. This program handles up to 45 variables and my program will operate with it, though the data must be transformed in batches of 13 variables or fewer. The cards can then be interspersed and the regression program will delete the excess dependent variables.

Table 3 (continued)

UNTRANSFORMED

MEANS X(10) TO X(14)

6.02	6.50	6.37	.00

STANDARD DEVIATIONS X(10) TO X(14)

3.410	2.926	3.031	.000

CORRELATION RATIOS X(10) TO X(14)

.40339	.32751	.40816	.00000
.53095	.57520	.52531	.00000
.32465	.34596	.41946	.00000
.09138	.14175	.18666	.00000
.16682	.14527	.21527	.00000
.53154	.52360	.61559	.00000
.16812	.26854	.24283	.00000
.16145	.22931	.17619	.00000
.29818	.44937	.37702	.00000
1.00000	.82901	.77535	.00000
.82901	1.00000	.89549	.00000
.77535	.89549	1.00000	.00000

TRANSFORMED

MEANS X(10) TO X(14)

490.82	490.82	490.82	.00

STANDARD DEVIATIONS X(10) TO X(14)

41.184	45.424	38.075	.000

CORRELATION RATIOS X(10) TO X(14)

.48075	.52919	.44448	.00000
.44141	.50027	.45037	.00000
.27214	.44100	.43205	.00000
-.00504	.23305	.25354	.00000
.08766	.40633	.32262	.00000
.56899	.59387	.67703	.00000
.35363	.41280	.31319	.00000
.21551	.27257	.13733	.00000
.30669	-.02732	.12141	.00000
.99999	.45266	.50732	.00000
.45266	.99999	.75240	.00000
.50732	.75240	.99999	.00000

relationships are much stronger.

In the case of the multiple R squared in Tables 4 and 5, the increase in variance accounted for between the untransformed and the transformed data was a modest 7 per cent, but the important result is to be seen in the plotbacks. These data were analyzed to see if the two failing scores of less than 350, runs number 0009 and 0021, could be identified from the data. In the untransformed plotback, neither case showed a value of less than 350; in the transformed data, the regression surface was in effect bent downward and identified both failing scores! This is an example of the primary reason for development of this procedure, the failure of linear regression equations to give a useful result even when the multiple correlation is fairly good. This bending of the regression surface and this correlation of curved data through transformations can be interpreted as a method of multiple curvilinear correlation and regression analysis.

In making predictions of future performance from new data, we use a program which reads in the transformation constants as well as the regression weights. The program then transforms the new data and computes the predicted values through the linear regression equation.

Sources of Error.

For data with a good range of values in all the variables, my experience is that the cubic polynomial transformations always bring some improvement in the correlations, and where the improvement is slight, this fact is taken as a verification that the data are already linear. With data that are dichotomized in the dependent variable only, roundoff error in solving the normal equations for the cubic polynomial will cause an occasional drop in a correlation. If there are several variables involved, however, the overall gain in variance accounted for as shown by the multiple R usually more than offsets the loss in variance for the drop in one variable or so. I have found that a quadratic least squares transformation will reverse this drop, but the quadratic transformation is generally not nearly as good as the cubic one and does not give as good a result in the multiple R as does the cubic.

If the data are dichotomized or trichotomized in both variables, the failure of the cubic or even the quadratic transformation becomes rather frequent, about one in five, due to the nearly vertical slopes of the relationships. Attempted solution of the normal equations simply produces overflow in the computer. Even so, the procedure could still be worth while if one is willing to go to the extra trouble to transform some variables and not others.

Another source of error is an insufficient number of significant digits in the input data. Experience indicates that a minimum of 4 significant digits in input is necessary to prevent excessive roundoff error in the solution of the normal equations even when a good range of values is present in all variables.

Final Suggestions.

It is suggested that at larger conputer installations a program might be written which could compute, try out, and select the best polynomial transformation up through the fifth degree for each variable. For smaller installations where it is practical to compute only one transformation, the cubic seems best, though the quintic is probably worth investigating.

Data straightening should also be a useful technique for factor analysis or any other procedure in which linearity is

Table 4

UNTRANSFORMED DATA

X	CORR. X	COEF.	F	ANALYSIS OF Y 01
01	.578075	-.012523	.001161	-4912523320
02	.781232	2.510669	.705612	5125106691
03	.398949	1.192242	.501091	5111922417
04	.685206	1.580214	.377998	5115802139
05	.619802	.665173	.861981	5066517270
06	.537416	.073669	1.127223	4973668636
07	.509511	.503662	.679341	5050366237
08	.273102	-.021606	.223086	-4921606009
09	.733979	1.198049	4.520080	5111980486
10	.877209	-.843564	.761309	-5084356360
11	.853182	-.080713	.008943	-4980713300

CONSTANT	MULTIPLE F	DF1	DF2	R SQUARE	RESIDUAL DELETE X 01
283.999910	4.051535	11	22	.669505	3341.038600

	RUN	OBSERVATION	PREDICTION	STD. DEV.	RESIDUAL	STD. RESID.	DISTANCE
	0001	551.000000	551.404460	36.149042	-.404460	-.005933	2.739006
	0002	524.000000	483.702110	34.276004	40.297890	.599668	2.794094
	0003	478.000000	503.300250	42.820373	-25.300250	-.351711	3.110948
	0004	442.000000	447.000990	29.234528	-5.000990	-.077207	2.852847
	0005	430.000000	504.758160	26.660867	-74.758160	-1.174444	2.280921
	0006	358.000000	481.233950	30.047650	-123.233950	-1.891681	1.911994
	0007	615.000000	612.520140	26.863120	2.479860	.038906	3.580497
	0008	551.000000	519.747500	32.043551	31.252500	.472881	1.580829
X	0009	325.000000	384.490350	35.176480	-59.490350	-.879202	5.079410
	0010	478.000000	480.231780	38.130501	-2.231780	-.032230	3.148011
	0011	583.000000	567.482420	29.314643	15.517580	.239430	3.591775
	0012	486.000000	441.587820	41.486726	44.412180	.624213	2.932177
	0013	559.000000	521.747790	31.118325	37.252210	.567472	2.210619
	0014	358.000000	411.741910	36.094089	-53.741910	-.788634	3.817078
	0015	551.000000	515.960900	34.300491	35.039100	.521316	2.665433
	0016	494.000000	493.658270	39.604338	.341730	.004877	3.711294
	0017	494.000000	538.129510	41.072772	-44.129510	-.622345	3.498808
	0018	583.000000	571.587380	33.846366	11.412620	.170383	3.178627
	0019	671.000000	679.940290	36.648075	-8.940290	-.130628	5.390885
	0020	559.000000	577.766080	43.950047	-18.766080	-.258440	3.954774
X	0021	293.000000	357.943710	29.611899	-64.943710	-.999974	4.716603
	0022	575.000000	574.408930	22.270449	.591070	.009542	3.121985
	0023	478.000000	542.179070	34.453822	-64.179070	-.953751	2.297220
	0024	478.000000	483.099270	40.537768	-5.099270	-.072228	3.759864
	0025	462.000000	453.580810	35.996065	8.419190	.123641	2.141199
	0026	559.000000	503.034770	31.435766	55.965230	.850574	2.494318
	0027	494.000000	475.560080	32.707736	18.439920	.277651	3.055932
	0028	462.000000	418.309070	31.245936	43.690930	.664941	3.787285
	0029	518.000000	498.499880	28.594887	19.500120	.302384	2.508432
	0030	551.000000	525.052130	41.376721	25.947870	.365026	2.086288
	0031	542.000000	552.577290	24.517204	-10.577290	-.168465	2.811941
	0032	414.000000	438.396490	38.283864	-24.396490	-.351888	3.142095
	0033	542.000000	468.848080	32.201547	73.151920	1.105577	3.317996
	0034	526.000000	404.518370	31.602096	121.481630	1.844078	3.971947

assumed. It should be possible to transform periodic data for use with linear models if good fitting periodic functions could be found.

Table 5

TRANSFORMED DATA

X	CORR. X	COEF.	F	ANALYSIS OF Y 01
01	.646592	-.037066	.011994	-4937066440
02	.777755	.164822	.249443	5016482155
03	.423603	.228222	.741834	5022822168
04	.748388	.532227	2.461307	5053222741
05	.622935	.133794	.216107	5013379399
06	.541339	.121212	.224274	5012121190
07	.419874	.303699	1.394496	5030369931
08	.298092	.431668	1.522364	5043166782
09	.611995	.781286	5.189080	5078128566
10	.787312	-.560829	1.033070	-5056082920
11	.732945	.244364	.245274	5024436400

CONSTANT	MULTIPLE F	DF1	DF2	R SQUARE	RESIDUAL DELETE X 01
-671.066220	5.763077	11	22	.742370	2604.432500

RUN	OBSERVATION	PREDICTION	STD. DEV.	RESIDUAL	STD. RESID.	DISTANCE
0001	551.000000	560.813750	29.408789	-9.813750	-.166615	2.736663
0002	524.000000	482.486450	26.367323	41.513550	.722695	2.762941
0003	478.000000	483.700250	27.307090	-5.700250	-.098484	2.219128
0004	442.000000	473.662180	26.566337	-31.662180	-.550318	2.324063
0005	430.000000	500.938310	20.683683	-70.938310	-1.288245	1.619861
0006	358.000000	497.184380	14.864574	-139.184380	-2.618493	.905418
0007	615.000000	589.904640	30.391661	25.095360	.422497	3.689797
0008	551.000000	479.927740	29.454460	71.072260	1.206175	1.499293
0009	325.000000	312.219230	40.884272	12.780770	.195452	6.907534
0010	478.000000	487.834180	32.306074	-9.834180	-.162819	2.434630
0011	583.000000	598.067880	26.532931	-15.067880	-.261964	2.863846
0012	486.000000	492.950060	31.539727	-6.950060	-.115847	2.396556
0013	559.000000	516.752060	27.933181	42.247940	.726183	2.029051
0014	358.000000	384.866210	37.489995	-26.866210	-.424266	4.288790
0015	551.000000	518.178960	33.536180	32.821040	.537465	2.894317
0016	494.000000	509.431340	31.701299	-15.431340	-.256854	3.480064
0017	494.000000	504.368400	33.202375	-10.368400	-.170298	3.144112
0018	583.000000	542.701720	36.356401	40.298280	.643131	3.678559
0019	671.000000	688.300020	39.490608	-17.300020	-.268098	6.010764
0020	559.000000	549.329320	30.414000	9.670680	.162781	3.194166
0021	293.000000	338.005280	35.328752	-45.005280	-.725085	5.903417
0022	575.000000	560.741220	14.674404	14.258780	.268519	1.949339
0023	478.000000	508.609380	28.237626	-30.609380	-.524808	2.225698
0024	478.000000	461.968510	37.025267	16.031490	.254266	3.431566
0025	462.000000	461.837240	39.484883	.162760	.002522	2.813478
0026	559.000000	517.620660	21.727444	41.379340	.746026	2.096788
0027	494.000000	529.011980	26.287368	-35.011980	-.609900	2.286177
0028	462.000000	425.764580	36.179148	36.235420	.579240	4.247105
0029	518.000000	498.930340	21.227510	19.069660	.345012	1.824780
0030	551.000000	539.700770	26.517619	11.299230	.196468	1.718618
0031	542.000000	561.186260	17.839140	-19.186260	-.354896	2.151836
0032	414.000000	444.402910	28.158670	-30.402910	-.521609	2.753899
0033	542.000000	527.050280	35.559394	14.949720	.240347	3.827166
0034	526.000000	435.565320	31.148203	90.434680	1.512581	3.712563

XIII

CONTRIBUTED PAPERS III

Chairman, PAUL AHMED, National Center for Health Statistics

NORTH CAROLINA SURVEY OF RECENTLY MARRIED PERSONS*: STUDY DESIGN AND PRETEST RESULTS

H. Bradley Wells, Elizabeth J. Coulter and Linda Wienir, University of North Carolina;
Monroe G. Sirken, National Center for Health Statistics.

I. Introduction

National statistics derived from marriage records are more limited than those based on birth and death records and this study is being done to test the feasibility of followback survey methodology for expanding national marriage statistics.

The North Carolina study is specifically designed to investigate the completeness and quality of information collected in followback surveys linked to marriage records. Three types of measurement error will be studied:

1) Response rates, including response after three consecutive mail questionnaires and personal interviews are attempted with nonrespondents to the mail survey.

2) Adequacy of response based on the completeness of information reported on the self-enumeration mail questionnaire, and

3) Accuracy and consistency of response based on comparison of information reported in personal interview for a subsample of respondents to the mail survey.

The population to be studied consists of all brides whose marriage license is filed with the North Carolina State Board of Health during the calendar year 1968 and the first four months of 1969. The study is being done in two phases (1) the pretest involving 192 brides and (2) the feasibility study which will include about 4300 brides.

The pretest was started in March and completed in June 1968. It was designed to test the acceptance and understanding of the pretest mail questionnaires both by mail response as well as by follow-up interviews.

The main purpose of this paper is to give a brief case history of the pretest phase. Some of the results from the stratified sample of 192 brides will also be described. Subsequent sections are:

II. Study Design
III. Response Rates
IV. Completeness and Quality of Responses
V. Revised Questionnaires
VI. Discussion.

Followback surveys linked to natality and mortality records have been undertaken since 1960 in a continuing program by the National Center for Health Statistics to expand and improve national vital statistics. Vital record files are the frames for these followback surveys and contacts for further information are sources whose identity and mailing address(es) are reported on the records. Contacting these sources by mail or combinations of mail plus telephone and/or personal interview have been successful in collecting additional data associated with recorded births and deaths, [1, 2, 3, 4, 5, 6]. In using marriage records as the base for followback surveys, the bride and/or groom will be the primary source of information. Conceivably other persons such as the issuing official, the officiant and if shown on the license, as in North Carolina, the parents of the bride or groom are potential secondary sources of data. Pratt[7] conducted a more limited study using marriage records as a base to determine the prevalence of pre-marital pregnancy. There are numerous examples of followback studies in other fields but these will not be considered here.

II. Pretest Study Design

A. North Carolina Marriage Registration

The Register of Deeds for each of the 100 counties in North Carolina is required to submit all marriage licenses filed with him during a month to the Public Health Statistics Section of the State Board of Health by the tenth of the following month. Processing of records in the State Board of Health is started five days later and by the fifth of the next month punched cards for these marriages are ready. Most marriages (about 95 per cent) are filed within the prescribed time limits. Thus, there is an average delay of one month between occurrence and filing of the license and a two month delay before the punched card is available.

In 1967, about 45,000 marriages were recorded in North Carolina. About one in five were of non-whites. In 20 per cent of the marriages the bride and groom were both non-residents of North Carolina. In 85 per cent neither had been previously married. Some 7,800 marriages were recorded for the five contiguous counties in the center of the State, which were included in the pretest.** About 37,000 occurred in the remaining 95 counties which were not included in the pretest but will be covered in the feasibility study.

*This study is being conducted under contract Number PH 43-67-764 with the National Center for Health Statistics and with the cooperation of the North Carolina State Board of Health.

**Alamance, Durham, Guilford, Orange and Wake counties.

B. Pretest - Sample Design

In order to minimize interviewer travel time and increase chances of locating nonrespondents for interview, pretest sampling was restricted to marriages in which the bride was a resident of North Carolina and the marriage had been performed during the months of July-August and October-November, 1967, in one of the five central counties, described above.

Pretest study population marriages were stratified on characteristics of the bride as recorded on the license as follows:

1) Race: a) white, b) non-white
2) Age (years): a) under 20, b) 20-29, c) 30 and over
3) Time between marriage (month of marriage) and mail survey: a) five or six months (October and November), b) eight or nine months (July and August)
4) Previous marital status: a) never married, b) previously married.

The top panel of Table 1 shows the number available for sampling in each cell. Notice the limited number in the under 20 year previously married, particularly the nonwhites, and in the strata for those 30 years and over never married.

Distribution of the sample is shown in the second panel of Table 1. The pretest design specified selection of eight marriages from each of the 24 strata and random allocation of two of these eight within each stratum to each of the four versions of the schedule being tested. Because of limited numbers in some strata it was not possible to balance on age. When a population stratum was exhausted, the remaining numbers were selected from the next older age stratum, keeping all other variables constant. Except for age the sample is balanced on the four demographic variables and on the four versions of the schedule.

C. Development of Pretest Questionnaires

Existing self-enumeration and interview schedules used by the National Center for Health Statistics and the Census Bureau, as well as schedules from a number of fertility surveys, were reviewed for content, definitions and possible wording of questions. Fairly rough formulations of possible questions were prepared. Several meetings of National Center for Health Statistics and University of North Carolina staff members were required before deciding on content areas and a reasonable number of items for self-enumeration schedules. Since one of the objectives of the feasibility study is to test the effect of content on response rates, it was decided to use three versions of the schedule in the pretest. The first included only basic demographic data on the couple and their parents, (Version D). The second consisted of Version D plus a section on family planning while the third included Version D plus migration and health sections. In order to partially test the format and wording of questions, a tersely worded form of Version D (Version A) was also developed. All questions were designed to be answered by the bride (wife). The first page of each questionnaire was a cover letter signed by the Director of the National Center for Health Statistics on Public Health Service letterhead. For each version these were identical except for one paragraph describing content. A different cover letter was used for each round of mailing but the questionnaires were exactly the same for every round.

D. Procedures

1) Mail Procedures

Within each pair of sample brides allotted to a particular version the first schedule was mailed to the bride, using her maiden and married

Table 1. DEMOGRAPHIC CHARACTERISTICS OF POPULATION AND SAMPLE OF RECENT BRIDES, NORTH CAROLINA MARRIAGE SURVEY PRETEST, FIVE COUNTY AREA*, SPRING, 1968.

Age (years)	Population				Sample			
	Never Married		Previously Married		Never Married		Previously Married	
	White	Nonwhite	White	Nonwhite	White	Nonwhite	White	Nonwhite
Eight to Nine Months Since Marriage								
Total	769	234	177	39	24	24	24	24
Under 20	348	103	4	--	8	8	4	--
20-29	403	115	75	14	8	8	12	14
30 or more	18	16	98	25	8	8	8	10
Five to Six Months Since Marriage								
Total	484	174	174	31	24	24	24	24
Under 20	259	78	8	1	8	8	8	1
20-29	214	87	76	7	8	8	8	7
30 or more	11	9	90	23	8	8	8	16

*Alamance, Durham, Guilford, Orange, and Wake Counties.

name, at her address as shown on the license, e.g.

Mrs. Mary Jones Brown
Bride's address,

and the second was mailed to the bride using only her married name at the groom's address, e.g.

Mrs. Thomas Brown
Groom's address.

A stamped return addressed envelope was enclosed for reply.

Three mailing were sent:

1) First class mail to total sample,
2) Certified mail to all nonrespondents to 1),
3) First class mail to all nonrespondents to 1) and 2).

When a first mail letter bounced (was returned by the post office unclaimed or undelivered) from either address, another first mailing was sent immediately to the other spouse's address on the license. The same time interval was allowed for subsequent mailing to these bounces; hence the timing was not the same as for regular mailings.

2) Interview Procedures

Personal interviews of subsamples of nonrespondents as well as respondents to the mail survey were made in order to explore in some detail objections and difficulties experienced by the sample brides in understanding or completing the mail questionnaire. Because of this objective, an attempt was made to simulate partially the mail survey setting. A mail questionnaire was given to the interviewee and she was asked to read the questions in sequence and give verbal answers which were recorded by the interviewer on a separate questionnaire. Interviewers observed and recorded difficulties both with sequence and with content. They also probed where indicated to get details about difficulties.

Three negro females - one single and two married - two white male graduate assistants and two white married females were recruited as interviewers and given one day of intensive training in Chapel Hill. One of the authors (LW) supervised interviewing and did all of the telephone interviewing.

III. Response Rates

One hundred and eleven (58 per cent) of the total sample brides responded to the mail questionnaire. Five brides returned the mail questionnaire unanswered, refusing to cooperate and they were not contacted for personal interview. Probability samples of 45 of the other 76 mail nonrespondents and 38 of the 111 respondents were chosen for personal interview. Table 2 summarizes the results of follow-up by various methods.

Interview results for the 45 sampled nonrespondents were as follows:

Total sample	45
Interview completed	27
Not located	10
Moved out of county	5
Unable to interview because of illness, etc.	2

Only the mail response plus interviews of mail nonrespondents are used to estimate rates. For completeness Table 2 includes other details about additional tracing and interviewing, both personal and telephone, of mail respondents and nonrespondents not sampled. These are not discussed further here although some of the responses are included in the Tables in Section IV.

The time pattern of mail responses in relation to the query is shown in Table 3. The number of first queries (i.e. indicated by a Roman Numeral I on the questionnaire) returned is higher than any other. However, about 40 per cent of them were returned after the second query (certified mail). The larger number of second queries returned in the third week, 28, indicated that it is a successful way of stimulating response. However, this does not indicate how effective using first class mail alone for the second query would have been. In addition it is not possible to estimate from these data what response rates would have been to a single one time questionnaire.

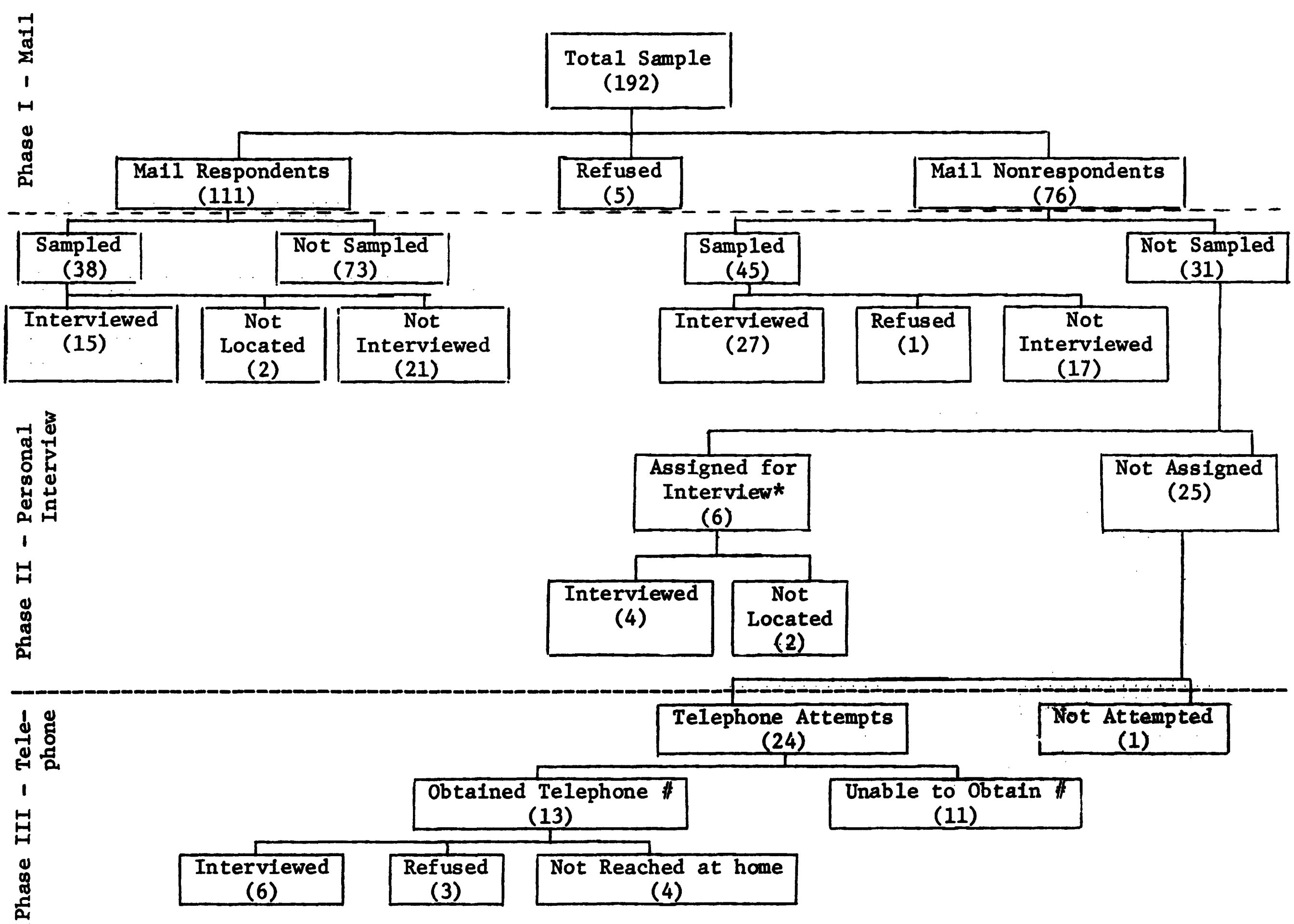
The cumulative response rate to the mail questionnaire was 57.8 per cent. The interview rate for mail nonrespondents was 60 per cent bringing the estimated total response rate to 81.5 per cent.

Table 4 shows total frequencies and estimated response rates for the four control variables from the marriage license as well as the different versions of the schedule and the address to which the first mail questionnaire was sent.

The most outstanding difference in total response rates is between marriages of 5-6 months duration, 90 per cent, versus 75 per cent for those of 8-9 months duration. The difference is mainly accounted for in the amount added by personal interview follow-up rather than total mail response. While this suggests that those married longer are more likely to have moved, this may also be related to other differences, (perhaps seasonal) associated with summer versus fall marriages. These differences will be explored in more detail in the feasibility study.

Mail response rates vary much more than total response rates. White rates are significantly higher than nonwhites. Women 20-29 responded at a higher rate than either younger or older age groups. Never married women responded at a somewhat higher rate than previously married women. Mail response rates for the two longer versions, B and C, were higher than for A and D but total response rates are about the same for all versions. Mail as well as total response rates are almost identical for bride's and groom's original mail address.

Table 2. SUMMARY OF FOLLOW-UP AFTER MAIL SURVEY (ALL VERSIONS) NORTH CAROLINA MARRIAGE SURVEY PRETEST, SPRING, 1968.



*Six (6) nonrespondents who were not in the sample were assigned to an interviewer in one county who completed other interviewing early.

Table 3. DISTRIBUTION OF MAIL QUESTIONNAIRES RETURNED BY QUERY AND WEEK RECEIVED, NORTH CAROLINA MARRIAGE SURVEY PRETEST, SPRING, 1968.

Sequence of Query	Total		Week Received						
			1	2	3	4	5	6	7 or later
	Number	Per Cent							
Total	111	100.0	13.5	18.9	36.0	12.6	13.5	4.5	0.9
First Query	58	100.0	25.9	36.2	20.7	8.6	8.6	-	-
Second Query	42	100.0	-	-	66.7	21.4	4.8	4.8	2.4
Third Query	11	100.0	-	-	-	-	72.7	27.3	-

Table 4. RESPONSE RATES BY DEMOGRAPHIC CHARACTERISTIC OF BRIDE OR QUERY NORTH CAROLINA MARRIAGE SURVEY PRETEST, SPRING, 1968.

Characteristics of Bride or Query	Total Number	Cumulative Response Rates: Mail Respondents Within Specified Weeks 1-2	3-4	5-7	Mail and* Interview	Non-** responses	Refusals
Total All Groups	192	18.8	46.9	57.8	81.5	15.9	2.6
Race:							
White	96	20.8	52.1	62.5	81.9	12.9	5.2
Nonwhite	96	16.7	41.7	53.1	81.2	18.8	-
Age:							
Under 20	45	15.6	42.2	51.1	81.9	13.7	4.4
20-29	73	24.7	57.5	65.8	83.5	15.1	1.4
30 and over	74	14.9	39.2	54.1	79.1	18.2	2.7
Marital Status:							
Previously Married	96	17.7	44.8	56.3	80.7	18.3	1.0
Never Married	96	19.8	49.0	59.4	82.2	13.6	4.2
Months Between Marriage and First Query							
Five or Six	96	20.8	44.8	56.3	89.5	7.4	3.1
Eight or Nine	96	16.7	49.0	59.4	74.5	23.4	2.1
Version of the Questionnaire							
A	48	18.8	39.6	47.9	82.6	17.4	-
B	48	16.7	52.1	62.5	80.2	17.7	2.1
C	48	20.8	50.0	62.5	86.8	6.9	6.3
D	48	18.8	45.8	58.3	76.3	21.6	2.1
Address Used in First Query							
Bride	97	18.6	45.4	56.7	83.5	13.4	3.1
Groom	95	18.9	48.4	58.9	79.3	18.6	2.1

*Per cent added by personal interview = $100 \left(\frac{\text{total nonrespondents in mail subsample}}{\text{total mail subsample}} \times \frac{\text{nonrespondents interviewed}}{\text{nonrespondents sampled}}\right)$

**Calculated by subtraction.

One must recognize the inadequacy of single factor analysis of response rates. Because of small frequencies multiple variable analysis was not attempted in the pretest. Perhaps the most important variable in analyzing response rates is education of the bride and this is available in the pretest only for respondents. However, education was added to the North Carolina marriage license in 1968 and it will be included in the feasibility study.

IV. Completeness and Quality of Response

A major use of the pretest results was in making decisions about questionnaires to be used in the feasibility study. Therefore, special attention was given to questions with relatively large numbers of unknown or incomplete responses in the pretest. These results are examined briefly in relation to:

1) type of question
2) characteristics of respondents who gave incomplete responses.

It is possible to improve the completeness and quality of data.*(see footnote pg. 6) However, results in this section are based upon results of unqueried mail returns plus personal and telephone interviews.

Questions had no provisions for checking a "don't know" answer, but space was provided for writing in responses other than those given on the form e.g. "Other (Specify)______."

Since only 48 brides were sampled for each version of the questionnaire, it is difficult to present a detailed analysis of quality by version. However, Table 5 shows some of the items which exhibit a difference (or lack thereof) between versions of the questionnaire in the proportion of unreported responses. Version A seems to stand out in items concerned with relatives as having a higher proportion of unknowns than the other versions, but this difference between versions was not consistent for other items, e.g. religion of the parents and birthdate of the wife and husband.

The most common type of problem that appeared in the pretest was a relatively large number of blank or unknown responses to questions about a relative--the wife's husband, and his or her parents. In case of mail queries the wife

Table 5. PROPORTION OF UNKNOWN OR UNREPORTED RESPONSES TO SELECTED ITEMS IN QUESTIONNAIRE BY THE VERSION MAILED, NORTH CAROLINA MARRIAGE SURVEY, SPRING, 1968.

Item	Version of the Questionnaire			
	A--Basic (tersely worded)	B--D plus family planning	C--D plus migration and health	D--Basic (regular wording)
Total Number of Questionnaires	34	38	39	37
Number of younger brothers and sisters:				
Wife	20.6	10.5	2.6	2.7
Husband	17.6	2.6	-	2.7
Number of married brothers and sisters of husband	23.5	NA	NA	NA
Highest grade of school attended:				
Wife	8.8	-	2.6	8.1
Husband	26.5	5.3	5.1	13.5
Highest grade of school completed:				
Wife	17.6	2.6	2.6	2.7
Husband	17.6	5.3	7.7	13.5
Religious preference of husband	5.9	7.9	5.1	10.8
Religion[a]:				
Father of wife	24.2	20.0	12.8	23.5
Father of husband	30.3	28.6	28.2	26.5
Mother of wife	15.2	14.3	2.6	11.8
Mother of husband	21.2	22.9	17.9	23.5
Year of birth[a]:				
Wife	12.1	14.3	12.8	23.5
Husband	15.2	22.9	15.4	20.6

[a]This item was not used during telephone interviews, so the base number of each version changes slightly (i.e., A = 33, B = 35, C = 39, D = 34).

NA = not applicable

*Enough requering was done in the pretest to find that incompleteness can be decreased considerably in this way. Of the 111 mail survey respondents 84 omitted or gave inadequate answers to one or more items on the questionnaire. Details for items with the highest frequencies were discussed in Section IV. Because of the tight time schedule in the pretest, requeries were mailed to only 50 of the 84 and 32 (64 per cent) were returned. Return rates were highest for those including inadequate response to migration questions. Lowest return rates were for those which included questions about income. In view of this experience we estimate that sending requeries to all 84 respondents with inadequate responses and incorporating the results into this paper, would have reduced most of the unknown or don't know frequencies in Section IV by 50 to 70 per cent.

Table 6. PROPORTION OF UNKNOWN OR NOT REPORTED RESPONSES FOR SPECIFIED ITEMS (ALL VERSIONS) FOR MAIL AND INTERVIEW QUESTIONNAIRES, NORTH CAROLINA MARRIAGE SURVEY PRETEST, SPRING, 1968.

Item	Number of Respondents to Which the Item was Applicable		Proportion Unknown or Not Answered	
	Wife	Husband	Wife	Husband
Birthdate:				
Year	137	137	16.1	19.0
Month	137	137	15.3	16.8
Day	137	137	16.1	19.0
Number of younger brothers and sisters	144	144	4.9	9.0
Highest grade of school attended	144	144	4.9	12.5
Highest grade of school completed	144	144	6.3	11.1
Income	144	144	12.5	18.8
Source(s) of Income	108	143	33.3	35.0
Month of first marriage of previously married	74	57	10.8	36.8
Year of first marriage	74	57	10.8	35.1
Month first marriage ended	74	57	13.5	42.1
Year first marriage ended	74	57	12.2	33.3
Education of parents				
Father	137	137	29.9	42.3
Mother	137	137	23.4	43.8
Religion of parents				
Father	137	137	20.4	29.2
Mother	137	137	10.9	21.9
Date of death of deceased father	54	48	25.9	47.9
Date of death of deceased mother	25	40	20.1	40.0
Separation of parents	141	141	10.6	13.5
Year of separation of separated parents	20	16	50.0	25.0
Divorce of parents	37	36	51.4	41.7
Year of divorce of divorced parents	12	13	41.7	38.5

presumably had an opportunity to discuss the questionnaire with her husband or possibly parents, while interview responses were given in most cases without her husband or other relative participating. However, information about parents was incompletely reported in both mail and interview responses.

Various questions about the husband resulted in relatively high frequencies of blank or unknown responses as compared with the wife. These questions were about the husband's number of younger brothers and sisters, his education, date and duration of previous first marriage and residence prior to marriage (See Tables 6 and 7). The pattern is similar yet the difference is less for the question on the highest grade of school completed.

When the couple had separated because the marriage had not worked, the wife experienced special difficulty with the items about her estranged husband, and his parents. One mail respondent who was legally separated refused to cooperate for this reason and three interview respondents who were no longer living with their spouse because of marital problems also indicated their reluctance to discuss these particular items.

Four types of information about the parents of the wife or husband were poorly reported--education, religion, dates of death of deceased parents, and whether separated and year of separation of parents and whether divorced and

Table 7. PROPORTION OF UNKNOWN OR NOT ANSWERED MIGRATION ITEMS IN RESPONDENTS TO MIGRATION QUESTIONNAIRE (VERSION C) MAIL AND INTERVIEW QUESTIONNAIRES, NORTH CAROLINA MARRIAGE SURVEY PRETEST, SPRING, 1968.

Item	Number of Questionnaire	Proportion Unknown or Not Answered	
		Wife	Husband
State of Residence Before Marriage	39	-	20.5
Number of Months at Address Before Marriage	39	2.6	33.3
Number of Years at Address Before Marriage	39	2.6	33.3
Living Arrangements Before Marriage*	39	-	25.6

*Lived alone, with parents, or with others.

year of divorce of parents.

The proportion of unknown or unreported answers was somewhat less for income than for some of the questions related to parents of the couple. Sources of income had a relatively larger number of incomplete answers; some responded "None" (i.e. no sources) when income had been previously reported, some others skipped the item or responded "don't know".

The difference between the proportion of unknowns for the wife and the husband is especially marked for the migration questions related to residence prior to marriage where there is about a ten fold ratio of unknowns for the husband as compared with the wife (Table 7).

The proportion of unknowns is generally higher for interviewed mail nonrespondents as compared with mail respondents but the numbers interviewed are too small for meaningful comparisons.

The proportion of unknown or unreported education and religion is higher among nonwhite than among white respondents, (Table 8) and there is a striking increase in proportions of unknown or unreported results with the age of the respondents, (Table 9).

Table 10 shows unreported and unknown answers to questions on family planning in the questionnaire returned by the brides who responded to the family planning version of the questionnaire. While frequencies are high they are

Table 8. PROPORTION OF RESPONDENTS FOR WHOM EDUCATION OR RELIGION OF PARENTS WAS UNKNOWN OR NOT REPORTED BY COLOR OF RESPONDENT, NORTH CAROLINA MARRIAGE SURVEY PRETEST, SPRING, 1968.

Parents		Mail Respondents		Interview Respondents	
		White	Nonwhite	White	Nonwhite
Number of Questionnaires Returned		60	51	11[a]	15
		Education			
Father of:	Bride	21.7	31.4	36.4	53.3
	Groom	30.0	45.1	54.5	73.3
Mother of:	Bride	25.0	19.6	18.2	33.3
	Groom	28.3	47.1	72.7	73.3
		Religion			
Father of:	Bride	10.0	31.4	18.2	26.7
	Groom	18.3	33.3	45.5	46.7
Mother of:	Bride	3.3	17.6	9.1	20.0
	Groom	11.7	21.1	18.2	40.0

[a]One interviewed white respondent was not asked these questions and is excluded from this table and the next one.

Table 9. PROPORTION OF RESPONDENTS FOR WHOM EDUCATION OR RELIGION OF PARENTS WAS UNKNOWN OR NOT REPORTED BY AGE OF RESPONDENT, NORTH CAROLINA MARRIAGE SURVEY PRETEST, SPRING, 1968.

Parents		Mail Respondents			Interview Respondents		
		<20	20-29	30+	<20	20-29	30+
Total Number of Responses		23	48	40	8	7	11
				Education			
Father of:	Bride	17.4	16.7	42.5	12.5	57.1	63.6
	Groom	26.1	29.2	52.5	37.5	71.4	81.8
Mother of:	Bride	8.7	14.6	40.0	-	14.3	54.5
	Groom	26.1	27.1	55.0	62.5	57.1	90.9
				Religion			
Father of:	Bride	21.7	10.4	30.0	12.5	42.9	18.2
	Groom	17.4	18.8	37.5	37.5	42.9	54.5
Mother of:	Bride	8.7	6.3	15.0	-	14.3	27.3
	Groom	8.7	12.5	35.0	25.0	14.3	45.5

Table 10. PROPORTION OF RESPONDENTS FOR WHOM SPECIFIED FAMILY PLANNING QUESTIONS IN MAIL AND INTERVIEW SURVEYS WAS SPECIFIED AS UNKNOWN OR NOT ANSWERED, NORTH CAROLINA MARRIAGE SURVEY PRETEST, SPRING, 1968.

Question	Number for Whom Requested	Proportion of Unknown Responses
Ever Thought About Number of Children?	35	17.1
Number Children Desired by Wife?	24	20.8
Number Children Desired by Husband?	24	20.8
Number Children Actually Expected by Couple?	26	30.8
Year Next Child Expected?	16	43.7
Prior Use of Contraceptives?	35	14.3
Type of Contraceptive Used?	19	26.3
Use of Contraceptives Expected if Not Used Previously?	21	38.1

not quite as great as for some questions on parents of the couple. Items which had the highest proportion of unknowns in this section were those about projections of future events, such as the number of children the wife thinks the couple will actually have, the year the next child is expected, and future use of contraceptives.

V. Revised Questionnaires

Three basic changes have been made for the feasibility study:

1. Version A, the tersely worded basic version, has been discarded--mainly on the basis of a lower mail response rate and a higher frequency of no response or unknown answers to individual questions. Thus only three versions will be used in the feasibility study.

2. The sections for the parents of the bride and groom have been omitted. This was done because of the high frequency of unusable answers to returned questionnaires and the responses to personal interviews.

3. The section on migration, formerly Part V of Version C, has been incorporated into the basic questionnaire and hence will be asked of all brides.

Other minor changes have also been made. All of these changes have resulted in a basic questionnaire for the feasibility study about one page shorter than the one used in the pretest. This will be called Version A and two other versions will be used: Version B will be Version A plus the section on family planning; Version C will be Version A plus a section on medical care.

VI. Summary and Comments

The objective of the pretest was limited primarily to developing a better mail questionnaire. Some of the specific pretest results, many of which could be predicted from previous experience in mail surveys, can be summarized as follows:

1. The second mailing stimulated response rates, perhaps because of certified mail.

2. Higher socio-economic groups apparently responded more completely and accurately than others.

3. Responses to questions on income were very incomplete (perhaps if such questions had been omitted, response rates would have been higher).

4. Usually more complete answers were given for questions about the respondent than for those about relatives.

5. Younger brides provided better quality responses than older brides.

6. Replies from mail survey respondents were more complete than those of nonrespondents to mail survey who were subsequently interviewed.

7. The tersely worded version of the basic questionnaire appeared to be less satisfactory than the less abbreviated version covering the same content.

More generally it was concluded that the pretest results were sufficiently encouraging to proceed with the feasibility study. Total response rates in the pretest were about 82 per cent and elimination of the questions about parents of the bride and groom should substantially increase response rates. Furthermore a systematic edit and requery system will reduce the relative high rates of missing or unknown answers in returned questionnaires.

Much more information was collected in the pretest than was needed to estimate response rates from combined mail and personal interview follow-up as indicated earlier in Table 2. Had a combination of mail, telephone, and personal interview in that sequence been used response rates might have been different. With adequate cost data on telephone personal interview and mail questionnaires and more precise estimates of response rates more efficient designs, perhaps including all three methods of follow-up, can be developed.

It is possible to study the effect of questionnaire content on response rates only in a limited way with the design used. Because of the unlimited number of possible kinds and combinations of content, decisions for this study were based on prior experience and the priorities for additional data about recently married couples. The influence of the sequential arrangement of content on response rates have not been tested in this study. For example income questions come near the beginning of the questionnaire in all versions.

There are numerous other factors such as color of paper, method of addressing the envelop, auspices, stamped versus prepaid return envelopes, method and size of printing, as well as the number and frequency of follow-up mailings and the type of mail e.g. certified versus first class, which might also be included in more elaborate study designs to estimate the impact on response rates. To study so many factors adequately at once will reuqire much more sophisticated sample designs.

References

[1] Sirken, M.G., Pifer, J.W., Brown, M.L., "Design of Surveys Linked to Death Records." National Center for Health Statistics, U.S. Department of Health Education and Welfare, September, 1962.

[2] Sutton, Gordon F., "Hospitalization in the Last Year of Life, United States, 1961." Vital and Health Statistics. National Center for Health Statistics, PHS Pub. No. 1000 - Series 22 - No. 1. Public Health Service, Washington, D.C. U.S. Government Printing Office, September, 1965.

[3] Wunderlich, Gooloo S. and Sutton, Gordon F., "Episodes and Duration of Hospitalization in the Last Year of Life, United States, 1961." Vital and Health Statistics. National Center for Health Statistics. PHS Pub. No. 1000- Series 22 - No. 2. Public Health Service, Washington, U.S. Government Printing Office, June, 1966.

[4] Wunderlich, Gooloo S. "Methods and Response Characteristics, National Natality Survey, United States, 1963." Vital and Health Statistics. National Center for Health Statistics. PHS Pub. No. 1000 - Series 22 No. 3. Public Health Service, Washington.

U.S Government Printing Office, September, 1966.

[5] Kovar, Mary Grace, "Visits for Medical and Dental Care, During the Year Preceding Childbirth, United States, 1963 Births." Vital and Health Statistics. PHS Pub. No. 1000 - Series 22 - No. 4. Public Health Service. Washington. U.S. Government Printing Office, May, 1968.

[6] Cappetta, Marlene and Nelson, Arne B., "Selected Findings of the National Natality Survey Pretest." North Carolina, 1967, unpublished report, January, 1968.

[7] Pratt, William Fares, A Study of Marriages Involving Premarital Pregnancies. Microfilmed dissertation. Ann Arbor, Michigan University Microfilms, 1965.

THE SYSTEMATIC BIAS EFFECTS OF INCOMPLETE RESPONSES

W. H. Williams, Bell Telephone Laboratories and U.S. Bureau of the Census

I. INTRODUCTION

Rotation sampling schemes are used for continuing studies in which there is interest in the estimation of change from month-to-month (say) as well as in separate estimates for the individual months. These rotation designs involve the month-to-month retention of some sampling units and the replacement of others. The details of rotation sampling 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 Grahm[5], and Kish[8]. It is to the point, however, to describe some applications.

A study of Bell System customers in western United States used a monthly sample consisting of three separate rotation groups. Each month one group appeared in sample for the first time, another for the second time, and the third had been in the two previous months. After three months in the sample each rotation group was dropped and did not reappear. The duration of study was eighteen months.

In the Current Population Survey (CPS), [6], conducted monthly by the U. S. Bureau of the Census one-eighth of the sample is new each month. Each new one-eighth group is retained in the sample for four consecutive months. It is then dropped for the next eight months, after which it is brought back into the sample again for four consecutive months. In this way each rotation group appears in the sample for a total of eight months.

In rotation group studies systematic biases have been observed in practice and the following examples appear to be typical.

In the Bell System study the average number of children per family for rotation groups appearing in the sample for the first time was 3.2. For rotation groups appearing in the sample for the second and third times the averages were 2.5 and 2.4 respectively. The average within rotation group variance of the monthly estimates was 0.1. Consequently, it appears that the first month may be significantly different from the second and third. How does one explain this apparent falling off in the number of children per household? Is it a systematic bias introduced by the interviewer or respondent? Or can a characteristic of the survey design or its implementation be responsible?

A similar characteristic appears in the CPS survey. Table 1 shows unemployment versus number of times in the survey for the CPS study. The data are taken from Waksberg and Pearl[15]. Unemployment appears to be higher for units which appear in the sample for the first and fifth times. [Recall that there is an eight-month lapse between the fourth and fifth interviews.] Why do these two peaks appear?*

Does the interviewer influence the respondent in such a way that he gives different responses from one month to the next? Such an hypothesis may be acceptable for the unemployment estimates but seems less likely for the number of children in the Bell System study. Similar behaviour exists for other characteristics in the CPS study, for example for estimated vacancy rates and families with salaries over $15,000, see Waksberg and Pearl[15].

Appearance in Sample

	1	2	3	4	5	6	7	8
Index	107.3	100.3	100.3	98.9	100.7	99.6	96.6	95.0

Table 1. CPS TOTAL UNEMPLOYMENT 1955-61

[INDEX NUMBERS,ALL GROUPS COMBINED EQUALS 100]

Before leaving this description of the problem, it is relevant to introduce "one-time" surveys which involve call-backs. These will be compared with rotation samples in the next section but it is to the point to present some data from one now. The data were taken from the paper by Finkner[7] and are presented in Table 2 below. Actually, the Finkner study was a multiple mail survey, but it is interesting because a systematic behaviour similar to the rotation group bias appears. Experienced practioners will of course recognize that this behaviour is common in call-back and mail surveys. It will be pointed out in this paper that this behaviour and the rotation group bias can have a similar cause.

	Number of Growers	% of pop^n	Avg. # of Fruit Trees per Grower
Response to first mailing	300	10	456
Response to 2nd mailing	543	17	382
Response to 3rd mailing	434	14	340
Nonrespondents after 3rd mailing	1839	59	290
Total population	3116	100	329

Table 2. FINKNER DATA - MULTIPLE MAIL SURVEY

II. INCOMPLETE SAMPLES

Population surveys are frequently conducted in such a way that all of the persons in a randomly selected area are to be included in the sample; other schemes will specify a subsampling of these persons, say by selecting every k^{th} household on a block. The remarks to be made in this paper apply to both cases, but to simplify the discussion and the formulas, it is assumed that all persons in the selected area are to be drawn into the sample. For the same reason, the higher structure of the sampling design is ignored. No loss in generality will result.

The number of persons in the selected area is denoted N, which may be known or unknown in practice, but seems more often to be unknown. The sampling scheme specifies that N persons are to be interviewed at some point in time, but in practice they rarely all are. To be specific, the objective of the survey is to interview N individuals in an area in such a way that the probability of inclusion, p_i', equals one, $i = 1,2,\ldots,N$. In practice, however, these probabilities may well be less than one with the result that a sample of $n < N$ persons is obtained. The expected number of persons is $\Sigma_{i=1}^{N} p_i'$ which equals N if all $p_i = 1$ and is less than N otherwise.

It was stated earlier that the survey which uses call-backs to obtain estimates at a single point in time has characteristics similar to rotation sampling. These can be seen by looking at the first visit as the first appearance in the rotation sample. The second visit (first call-back) is the same as the second appearance in the rotation sample if those persons interviewed at the first visit are considered to be included at the second visit with probability one. [They are not actually visited twice but the data obtained are simply carried over.] A difference is that call-back surveys use the assumption that the characteristics under observation do not change with time, while rotation samples are designed to estimate this change.

In both call-back and rotation surveys, estimation difficulties arise because the probabilities with which a response is obtained are unknown. Estimation is usually carried out by assuming that these response probabilities are equal. What are the effects of this practice? For call-back sampling, the problem has long been recognized and papers have appeared on the subject. It seems unnecessary to trace these in this paper except to point out that a good description of the work has been given by Kish[8], pp. 532-62. The papers by Politz and Simmons[10],[11], and Hartley[12] are relevant to the work in this paper, in that an attempt is made to estimate the individual response probability.

In rotation sampling the effects of these unknown probabilities do not seem to have been discussed. An additional difficulty is that these probabilities are undoubtedly changing from one appearance in the survey to the next, and probably are doing it in a systematic way. This problem is discussed in the next section in such a way that the results are applicable to any design which involves periodic reinterviews.

III. THE EFFECTS OF THE UNKNOWN PROBABILITIES

3.1 At the First Appearance

Suppose that the N units in the sampled area have characteristics y_i, $i = 1,2,\ldots,N$. The objective is that responses be obtained from each of them with probability $p_i' = 1$, $i = 1,2,\ldots,N$. However, as pointed out earlier the interviewing method is not likely to be that successful and $p_i' = 1$ will <u>not</u> be achieved for all i units. Then the expected sample size (number of responses) is $n_1^* = E(n_1) = \Sigma_{i=1}^{N} p_i'$, where n_1 is the number of interviews actually obtained. Next, an estimate of the mean is formed as $\hat{Y}_1 = \left(\Sigma_{i=1}^{n_1} y_i\right)/n_1$, which is a ratio estimator with expectation, $E(\hat{Y}_1) \doteq \left(\Sigma_{i=1}^{N} p_i' y_i\right)/\left(\Sigma_{i=1}^{N} p_i'\right)$. This expectation is approximate but the technical bias of the ratio estimator is not important here.

The incomplete response has effectively introduced an additional level of sampling into the overall design. The effect on total variance is probably not large because this additional component of variance comes in at the lowest level in the sampling design. The bias effects may be quite another matter however, since the probabilities of inclusion at the last stage are unknown and may very well have a systematic behaviour.

3.2 Rotation Sampling and Call-Backs

The second time the selected persons are to be interviewed there can be little doubt that the probabilities of actual inclusion will have changed from the first interview. There are a number of reasons for this. One is that it would be expected that the information gained at the time of the first interview period, (T_1), would increase the probability of a response at the second, (T_2). The interview team probably knows the area and the availability characteristics of some of the individuals better at T_2 than at T_1. Consequently, a survey manager would naturally expect that the number of responses obtained would tend to go up at T_2. It seems unlikely, however, that <u>every</u> unit will have a larger probability at T_2, some could conceivably decline. The number of refusals for example typically increases the longer a group has been in the sample. Specifically, the units will have probabilities p_i'' associated with them at T_2 and many of these will be different from the p_i' at T_1. In rotation sampling it will also be expected that some of the characteristics y_i, $i = 1,2,\ldots,N$ will have changed. One of the purposes of rotation sampling is to obtain efficient estimates of this change. However, in this paper we wish to study the possible effects of the changes in probabilities and so to insure that there are no confounded factors, it is assumed that the y_i do <u>not</u> change from T_1 to T_2.

Given this hypothesis, rotation sampling and call-back surveys are very similar.

Consequently, with the above assumptions, $n_2^* = E(n_2) = \Sigma_{i=1}^N p_i''$, is the expected sample size, and the estimator, $\hat{\bar{Y}}_2 = \left(\Sigma_{i=1}^{n_2} y_i\right)/n_2$, has the approximate expectation $E(\hat{\bar{Y}}_2) \doteq \left(\Sigma_{i=1}^N p_i'' y_i\right) / \left(\Sigma_{i=1}^N p_i''\right)$.

3.3 The Special Case of Proportions

A case of special interest is that in which there are two classifications such as employed and unemployed. It should be emphasized that these two categories are referred to as employed and unemployed because this work was originally suggested by consideration of the characteristics of unemployment statistics. The extent to which these models actually apply to unemployment statistics has not yet been determined. Then if (i) p_e' denotes the probability of an employed person actually being interviewed at T_1, and (ii) p_u' denotes the analogous probability for an unemployed person, and (iii) y_i equals one if unemployed and equals zero if employed, then $\hat{R}_{u1} = n_{u1}/(n_{e1}+n_{u1})$ and $E(\hat{R}_{u1}) = N_u p_u' / \left(N_e p_e' + N_u p_u'\right)$. Similar expressions can be written down for the unemployment rates at T_2. The generalization to more categories presents no difficulties.

3.4 The Bias Effects of the Unknown and Changing Probabilities

Under the assumption of no changes in the characteristic y_i it would be hoped that the expectations of $\hat{\bar{Y}}_1$ and $\hat{\bar{Y}}_2$ would be the same and equal to $\bar{Y}$ the population mean. Is this true? And if not, what statements can be made?

The technical question being asked is how does $\left(\Sigma p_i'' y_i\right) / \left(\Sigma p_i''\right)$ compare with $\left(\Sigma p_i' y_i\right) / \left(\Sigma p_i'\right)$? To this end the following points can be easily made.

(1) If $p_i'' = kp_i'$, the expectations at T_1 and T_2 are the same.

(2) If the p_i''s are randomly associated with the y_i's, the expectation at T_1 is equal to $\bar{Y}$. Similarly, if the p_i'''s are randomly associated with the y_i's the expectation at T_2 is equal to $\bar{Y}$. Consequently there is no bias at T_1 or T_2 and no systematic change from T_1 to T_2.

(3) What happens under the more realistic assumption that the probabilities at T_1 are related to the characteristic y_i and that more information and experience on the part of the interviewers at T_2 brings these probabilities closer to equality and to one? To answer this, suppose that $p_i = ky_i^{\alpha}$, and that all the y_i's are positive. Then it can be easily shown that the estimator $\Sigma p_i y_i / \Sigma p_i$ increases monotonically with α.

(i) As a first example suppose that $p_i' \sim y_i$ and $p_i'' = 1.0$. This means that at T_1 the units with larger y values have a higher probability of entering the sample and that at T_2 all units enter the sample. The latter choice of all $p_i = 1$ is the survey manager's idealized goal and would be a result of an efficient interview program at T_2. Since $p_i' \sim y_i$ at T_1 corresponds to $\alpha = 1$ and $p_i'' \equiv 1$ at T_2 to $\alpha = 0$, it follows from above that $E\hat{\bar{Y}}_2 \leq E\hat{\bar{Y}}_1$, the equality occurring if all y_i's are equal. It is important to notice that this systematic change comes about solely as a result of changes in the probabilities and <u>will occur even though there has been no change in the characteristic being measured</u>.

(ii) As a second example, suppose that $p_i' \sim 1/y_i$ and $p_i'' = 1$, so that the larger units have a smaller chance of appearing in the sample at T_1. Then $E\hat{\bar{Y}}_2 \geq E\hat{\bar{Y}}_1$ and a systematic change appears in the opposite direction. This again is solely a result of changing probabilities because the y characteristics have been assumed to be constant in the time period from T_1 to T_2.

What can be said about the specific case of unemployment? First, it can be easily shown that $E(R_u) \lesseqgtr$ True Rate, $\Leftrightarrow p_u \lesseqgtr p_e$, which is an obvious intuitive result. Secondly, if the probabilities for employment and unemployment each change in different proportions from T_1 to T_2, as follows, $p_e'' = k_1 p_e'$, $p_u'' = k_2 p_u' = ck_1 p_u'$, then it can be easily shown that, $E(R_{u2}) \lesseqgtr E(R_{u1}) \Leftrightarrow c \lesseqgtr 1$. For example, if $c < 1$, $k_2 < k_1$, and $E(R_{u2}) < E(R_{u1})$. This means that if the biggest change in probability from T_1 to T_2 is associated with employed persons then you <u>must</u> get a decrease in the expected value of the estimator solely as a result of this change. In this situation, it would seem likely that at T_1, $p_u > p_e$, which in fact concurs with the field experience of Deming[13], Harris[14], and Kish[8].

In this case it is interesting to look at some numerical results. Suppose that $N = 10{,}000$, $N_u = 400$, so that $R_u = 0.04$, then simple calculations yield the figures in Table 3. If case i represents the situation at T_1, and an effort is made at T_2 to improve the response so that case ii describes the resultant situation then we see that there has been a five percent change in the expectation of the estimate with <u>no change in actual unemployment and in spite of a high response rate</u>. Case iii simply shows that without knowledge of the p's there is no way of knowing whether R_u is being over or underestimated. Case iv shows that a three percent bias is possible with probability differences which intuitively one would probably judge to be very small.

Cases iv and v are interesting to consider together. If at T_1, (case iv), p_u is slightly higher than p_e (as indicated), and if as a result of any "unobservable" characteristic of unemployed persons p_u drops, then a comparison of the cases shows a ten percent drop in $E(\hat{R}_u)$ with virtually no change in the response rate. In practice, however, the response rate does in fact improve from T_1 to T_2. If this response increase was a result of an increase in p_e, and if p_u was prevented from improving by a hard core of unobservable unemployed persons then cases vi and vii show what may happen. Specifically, there has been a five percent change in the expectation of the estimator. It is possible to construct examples like this indefinitely. To what extent any of these factors apply to a specific survey, each practioner will have to decide for himself.

Case	p_e	p_u	$E(\hat{R}_u)$	$E(n)$
i	0.90	0.95	0.0421	9,020
ii	0.95	0.95	0.0400	9,500
iii	0.95	0.90	0.0380	9,480
iv	0.95	0.98	0.0412	9,512
v	0.96	0.90	0.0376	9,576
vi	0.92	0.95	0.0413	9,192
vii	0.98	0.95	0.0388	9,788

Table 3. POSSIBLE UNEMPLOYMENT BIASES

3.5 Coverage

The case in which some p_i equal zero is usually referred to as a coverage problem. It means that some persons who should appear in the sample have no chance of actually entering. It follows from the earlier discussion that efforts to improve the coverage will contribute to the rotation bias effects by increasing some of the p_i. Unfortunately, if the group which is not being covered tends to have a certain characteristic the bias effects can be dramatic. For example, suppose that there is a hard core of "unobservables" who tend mostly to be unemployed. To be specific, consider the example of section 3.4 in which $N = 10{,}000$ $N_u = 400$ and $R_u = 0.040$. In addition assume that there has been a coverage loss of one half percent or fifty persons and that twenty percent of these are unemployed. Then with equal probabilities $p_u = p_e = 0.95$ it is easy to calculate that $E(n) = 9452.5$ and $E(\hat{R}_u) = 0.0392$ so that a two percent bias has been introduced. Next suppose that the "uncovered" group has even more unemployment than supposed, specifically that out of the fifty persons missed, twenty are unemployed. Then $E(n) = 9452.5$ as before, but $E(\hat{R}_u) = 0.0382$, a four and one half percent bias. To push the example still further, suppose that the coverage problem jumps to one percent with $p_u = p_e = 0.95$ and forth of the "unobservables" are unemployed. Then $E(n) = 9405$, $E(R_u) = 0.0364$, and the bias has jumped to nearly ten percent. Finally, if there is a one percent coverage error, forty of whom are unemployed, coupled with $p_e = 0.95$ and $p_u = 0.90$, then $E(n) = 9387$ and $E(R_u) = 0.0345$ which is a bias of about fourteen percent. Notice that the response rate is not necessarily indicative of the bias behaviour. In order of their presentation above, the values of $E(n)$ were 9452.5, 9452.5, 9405 9387 which for most practical considerations would be considered to be unchanged.

It will be recalled from section 3.4 that in the experience of a number of practioners $p_u > p_e$ at T_1, and it was shown that, if true, this would cause an upward bias. For example if $p_u = 0.95$ and $p_e = 0.90$ $E(\hat{R}_u) = 0.0421$. Consequently, if, (i) $p_u > p_e$ at T_1, (ii) $p_e > p_u$ at T_2, (iii) a coverage problem appears which is associated with unemployed persons, then combining the calculations made above shows that $E(R_u)$ may drop from 0.0421 to 0.0345. This is a change of twenty percent without any real change in unemployment. It is relevant that the data of Waksberg and Pearl[15] suggest that coverage tends to have a rotation group bias type behaviour. This has also been the Canadian experience[16].

IV. SUMMARY AND DISCUSSION

(1) In this paper it has been shown that systematic changes in the response probabilities can cause the type of systematic bias that has been observed in rotation sampling. Under certain assumptions, the expected value of the estimator must change from the first time to the second time that a rotation group appears in the sample.

(2) Are the basic hypotheses reasonable?

(i) The first necessary hypothesis is that the probability of a response actually being obtained is related monotonically to the characteristic exhibiting the bias. It seems clear from experience that this can actually occur. Indeed in the case of number of children per family, it would be surprising if it were otherwise. Surely the families with children are more likely to be found at home. The suggestion of such an association is not new. There is a large literature on this problem, see Kish's discussion, [8].

(ii) The second hypothesis required is that the probability of response changes from T_1 to T_2. In many studies there can be little doubt that this is true, because there is a systematic, significant increase in the response rate. Such a significant change in the response rate must be a result of changing probabilities. In particular, an increase in the response rate must mean that an overall increase in the response probabilities has occurred. This is not surprising because the managers of every survey are working towards the goal of maximization of the response rate. On the other hand, it is important to notice that there can be systematic biases without any noticeable change in the response rate.

(3) The hypothesis that the y_i do not change from T_1 to T_2 has also been used. This hypothesis was made in order to study the effect of changes in the p_i's in an otherwise static

situation. In a forthcoming paper by Mallows and Williams[17], this assumption has not been made and some very interesting results have been obtained. Three of these are as follows.

(i) The magnitude of a potential bias can be made very large, even in very innocent-looking situations. For example, biases of one hundred percent can be obtained even with response rates over ninety percent.

(ii) To attempt to study socio-economic patterns by the study of matched individuals may be highly misleading. For example, is it true that people tend to answer questions about their economic status differently in a first interview than a second? To examine this, it is tempting to construct estimates which are based only on those individuals who appear in the survey at both 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 is shown to be a false statement by Mallows and Williams[17].

Obviously, however, if there are systematic reporting changes, these changes will evidence themselves in the estimates. Such phenomena may or may not exist and this paper does not concern itself with their presen e or absence.

(iii) An estimation scheme has been developed and is describedin the paper.

The author wishes to convey his appreciation to David Brillingerand Colin Mallows for a number of helpful discussions.

V. LITERATURE CITED

[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, 492-509, 1964.

[6] 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.

[7] Finkner, A. L., Methods of sampling for estimating commercial peach production in North Carolina. North Carolina Agr. Exp. Stat. Tech. Bull. 91, 1950.

[8] Kish, Leslie, Survey Sampling, John Wiley and Sons, New York, 1965.

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

[10] Politz, A. N. and Simmons, W. R., An attempt to get the "not at homes" into the sample without call-backs. Jour. Amer. Stat. Assoc. 44, 9-31, 1949.

[11] Politz, A. N. and Simmons, W. R., An attempt to get the "not at homes" into the sample without call-backs. Jour. Amer. Assoc. 45, 136-137, 1950.

[12] Hartley, H. O., Discussion of a paper by F. Yates, Jour. Roy. Stat. Soc., 109, 37, 1946.

[13] Deming, W. Edwards, Verbal discussion, Consultant in Statistical Surveys, Washington, D.C., 1967.

[14] Harris, Louis, Verbal discussion, Louis Harris Associates, New York, New York, 1967.

[15] 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.

[16] Fellegi, Ivan, Verbal Discussion, 1967.

[17] Mallows, C. L.; and Williams, W. H., Systematic biases in panel surveys. To be submitted to the Journal of the American Statistical Association.

THE DEMAND FOR NEIGHBORHOOD MEDICAL CARE: STATISTICAL ANALYSIS

Irving Leveson and Milton Brafman*
The RAND Corporation

Introduction

Past studies of the demand for medical care have focused on medical care as a whole or on hospital inpatient care. However, ambulatory medical care has been growing rapidly. Indications are that it will continue to grow rapidly as we pursue policies intended to promote alternatives to expensive inpatient care. There is, in addition, a widespread belief that ambulatory medical care may have a particularly high payoff in improved health. Until recently, work on the geographic distribution of medical care has centered on supply and/or on large areas, except for case studies providing results in forms which are difficult to generalize. With heightened interest on health services for the poor and concern with reaching ghetto populations, analysis of health care for small areas takes on increased importance. The demand for ambulatory care for neighborhoods has been neglected in economic analyses. Programs will increasingly be judged in the future by how successfully they reach their target populations, and in order to improve outreach, demand studies are needed. Hopefully these studies will throw light on the demand for other social services as well.

Significant developments have taken place in thinking about demand determinants in recent years. Furthermore, there are important differences in the role of variables influencing a particular type of service at the neighborhood level from their effects on broader aggregates. The choice of medical services can be treated as an investment decision in which the costs of experimentation with new sources of care weighed against the expected benefits. Both private and social benefits are relevant for government decisions. Health status is best considered explicitly as a determinant of the demand for medical care, rather than being partially represented by other variables categorized as "tastes". With the growth of Medicaid and other forms of "insurance", the traditional demand determinants of income and money price lose importance. Great weight falls on the role of non-money "price" variables, which may have a great impact on accessibility of care, such as opportunity costs of the consumer's time, inconvenience due to factors such as limitations to physical mobility, and consumer information. At given levels of other variables, "quality of care" becomes a "price" variable also.

The statistical analysis considers an entire population - the residents of a housing project for the elderly in which a clinic was established. Determinants of whether or not people use the clinic are analyzed in Chi-Square tests and a dummy variable multiple regression analysis is performed. Some evidence from other clinics is also examined.

The availability of services does not necessarily mean that they will be used, even if properly tailored to the needs of the community. Information about the determinants of demand for medical care in facilities will help in predicting utilization and identifying needs, and hopefully also in devising ways to encourage utilization. Here we concentrate on the primary aspect of utilization -- whether or not services are used at all. One of the methods of organizing medical care, which has great potential for reaching the community and economizing on costly inpatient and nursing home care, is the provision of services in clinics established in housing projects for the elderly. One such unit, established with these points in mind is the Queenbridge Health Maintenance Service (QHMS), which was set up in November of 1961 by the New York City Departments of Health, Hospitals, Welfare and other agencies. The unit offered a variety of services, including home care without charge to the 1,400 residents of Queensbridge Housing Project in Long Island City, New York. Extensive efforts were made to inform residents of the on site availability of free services which previously required travel of four miles or more. An evaluation of the project has been conducted by Kuo _et. al._[1] Here the determinants of whether or not persons registered for the service are analyzed. Particular attention is paid to the role of health status in the demand for medical care, to separating the effects of income and education in disaggregated data, and to non-money "price" and information variables.

The Data and Methods

The sample consists of the 1,219 of the approximately 1,400 residents of the Queensbridge Housing Project for whom information has been obtained between January 1963 and the Fall of 1964. Only persons age 60 and over are included since

---

* Any views expressed in this paper are those of the authors. They should not be interpreted as reflecting the views of the RAND Corporation or the official opinion or policy of any of its governmental or private research sponsors.

The material in this paper was prepared while at the New York City Health Services Administration. It was incorporated in a report entitled Medical Care and Income, An Economic Analysis, published by the Gouverneur Economic Research Project. The effort was financed under Public Health Service Research Grant No. R01 CH00239 and New York City Health Research Council Contract U-1153, with additional support from the Carnegie Foundation. We would like to thank Dr. Paul Densen, Dr. Nicetas Kuo, Mrs. Ellen Jones and Jerry Weston for making materials available. Sid Binder contributed freely of his time in facilitating the data processing activities.

use of the clinic was limited to that age group. Of the total, 638 had registered (participants) with the QHMS and 581 had not. As a means of separating the effects of health status from other variables, a subsample of 343 persons who reported that they had heart conditions is examined. The group of heart patients is of particular interest in view of the emphasis of the Regional Medical Programs[2] and is large enough to make statistical analysis meaningful. Among the heart group, 201 persons participated and 142 did not. Information on non-registrants was collected in the Fall of 1963 in interviews conducted by the National Opinion Research Center, while registrant socioeconomic data was collected at various dates at the QHMS. The initial groups were interviewed at the time of registration. It is therefore possible that the information obtained is differently influenced by degree of satisfaction with the clinic or knowledge gained about one's health.

Variation in the percentage of persons participating in the clinic across socioeconomic groups is examined. In one-way classifications, for both the total and heart groups, we try to refute the null hypotheses that the percentage participating does not vary among socioeconomic groups by Chi-square "goodness of fit" tests. Specific hypotheses concerning the variables examined are introduced at that time and the effect of the control for health status is considered for each variable.

A multivariate analysis of the heart group is then performed using the dummy variable multiple regression technique. All variables take on values of zero or one. For example, the dependent variable is coded 1 for participants and zero for nonparticipants. For variables which cannot appropriately be dichotomized, a series of dummy variables replaces a single variable. Each variable in the set represents a class interval of the continuous form, and observations are coded 1 or 0 depending on whether or not the value of the variable falls in a given interval.[3] One variable is omitted from each set of dummies and becomes the base. The regression coefficient for any dummy in the set then indicates by how much each class interval of the independent variable differs from the base group in its value of the dependent variable.[4]

The chief disadvantage of the dummy variable technique is that it does not conform with the assumption of homogeneity of variance of the regression model. This still enables estimation of unbiased regression coefficients when other regression problems are absent, but casts doubt on the validity of the usual tests of statistical significance. For this reason, the analysis utilizes a Chi-square test of the significance of sets of dummy variables which depends only on the regression coefficients and not on the distribution of the regression residuals.[5] As in the one-way Chi-square tests, the "expected" values are the numbers of participants and non-participants that would be observed in each class interval of the independent variable if the average percentage participated in each class interval. Instead of comparing these with true actual values, however, pseudo-actual values are derived from the set of regression coefficients for a variable. These differ from the actual values in that other variables are held constant. It should be noted that in the cases in which the Chi-square test was compared with the F-test, nearly the same levels of significance were indicated.

The dummy variable technique has certain advantages over use of continuous forms. While it is especially suited to analysis of survey data, its use for this purpose has been extremely limited. Some of these advantages are great enough so it would pay at times to create sets of dummy variablesfrom continuous forms. For these reasons it is worth stating the advantage clearly. There is no need to make an assumption of linearity, and it is easy to fit functions which are nonmonotonic. By cross-classifying variables, tests for interaction can easily be made without the usual problems of high multicolinearity. Since all observations are used, it is possible to avoid many of the limitations of cross-classification. In effect, small sample rows and columns are pooled on the assumption that the effects of variables are additive except where specific interaction terms are introduced. Another advantage is the ability to retain information for persons with unknown values of certain variables but known values of others, by including an "unknown" category in the set of dummies. Since survey data often come in class intervals, measurement error associated with a choice of midpoints is avoided. Also, since errors in the data may be concentrated at certain values of a variable, this will show up as a high standard error for that coefficient, so that it may be possible to reject one or more coefficients without losing information on the remaining set of dummies or obtaining a biased coefficient for the entire variable.

The Results

Table 1 shows the percentage participating by level of each variable, for all residents and for those reporting heart trouble. Table 2 indicates the levels of statistical significance in the test of whether the percentages differ from equality among levels of any variable. The percentage participating is slightly higher in the heart population than for all residents, 58.6 percent compared to 52.3. This is expected since a group known to have any given health problem will tend to have poorer health than the overall population, some of whom have no health problem, and because of the existence of a health problem contributes to the decision to seek medical care. Similarly, among the total population, older persons are more likely to be participants because they will tend to have more health problems. However, among the heart group, age is not significant. This may be because aging does not produce a worsening of heart problems which requires more medical care, the care is not sought, or when conditions become more serious it is sought in places other than the clinic. Another way of looking at the effect of presence of a heart condition is to note that younger persons with heart conditions are more likely to use the clinic than others, but at older ages where a greater percentage of persons will have health problems requiring medical attention, the two groups participate about equally. A further indication of the role of health status is given by self-rating of health, which similarly is to age, is highly significant for the total population but not significant for heart population. Sex differences are small and not statistically significant.

The data for the total population showed greater participation for whites even though non-whites might be less able to afford alternative sources of care, and a substantial proportion of the clinic personnel were non-white. It was hypothesized that this too could be attributed to health status, since non-whites surviving to age sixty or more would have to be hardier relative to their initial cohort than whites who are likely to have had more medical care when needed and a better environment. Surviving non-whites would tend to have fewer health problems and, if this were the case, the color difference would not appear among the heart group. In fact, the color difference is very much smaller for the heart group and is not statistically significant.

Not all effects of health status are to increase the demand for medical care. Mobility limitations clearly reduced participation even though home visits could have been arranged. Again the effect of health status can be associated with the presence of a disabling condition since within the heart group mobility was not significant. Mobility status can be considered a kind of a price variable, indicative of the amount of inconvenience or discomfort of getting about, or of the costs that would have to be incurred to avoid that discomfort. Another "price" variable which was included in the analysis is travel time to the person's usual source of care prior to clinic registration. Participation was greater among persons having to travel an hour or more in both groups. However, differences were not statistically significant, perhaps because the variable is poorly measured. There is some indication in the results that health insurance coverage leads to greater use of alternative sources of care, as its effect on relative prices would suggest. To some extent, both health status and price effects probably account for the large differences in participation of persons classified according to the usual source of medical care they had prior to the availability of the clinic. It is disturbing that over one-third of the total population and over one-fourth of the heart group both did not participate and did not report any prior usual source of care.

It was hypothesized that the higher the income of a person, the better he could afford to purchase medical care from alternative sources if he believed it necessary, more appropriate, of higher quality or otherwise desirable. Since much of the income for the aged represents pension payments, we probably have a measure of lifetime as well as current income. Income turns out to be a very powerful variable for both the total and heart groups.

It was possible to test the effects of several variables reflecting various aspects of the availability of information to the project residents. Both marital status and number of people in the household were intended to reflect the number of instances a person heard about the clinic, while marital status might also have operated through concern for the health of one's mate. However, neither was significant. Employment status, on the other hand, intended to reflect the awareness of alternatives, and willingness and ability to operate in the outside community, did prove important. However, this may also reflect good health and less need for medical care. Both length of residence in the project and whether foreign born, intended to reflect knowledge of alternative sources of care, were statistically significant.

It was expected that persons with more formal education would be more likely to participate because of greater concern for the future, receptiveness to information, and other factors. However, the education variable was not found significant in the univariate tests. The dummy variable multiple regression analysis of the heart group generally produces results which closely mirror those already presented. One important change was expected, however, which did materialize. Education and income tend to be highly positively correlated with each other, so that education would tend to reflect the effects of income. A zero effect of education could be produced as the resultant of a negative effect of income on participation and a positive effect of education. Table 3 presents the regression results, and the levels of significance are given in the last column of Table 2. The adjusted regression coefficients, calculated by a method given by Melichar,[6] indicate the deviation of percent participating in each group from the overall mean. When income and other variables are held constant, education is seen to have a large positive effect on participation which is statistically significant.

In summary there is evidence that health status is a major determinant of whether people seek medical care. Differences in care can be discerned even when the patient himself provides broad diagnostic information. Furthermore, the tests suggest that important variation in medical care sought can be isolated by simply classifying persons as to the presence or absence of a condition, without information on its severity. The success of crude self-classification may be that such a concept influences the patient's attitudes which are relevant for behavior, though one's own evaluation may not be as well related to functional status as a physician's classification. Income appears to be a powerful determinant of patterns of care. The effects of price variables such as mobility status also appear to be important. Information variables differ greatly in their impact but as a whole seem to contribute substantially to the observed variation in use and sources of medical care. Most notably, the more educated and recent residents of the area appear more likely to use clinic services. When other variables were held constant, color was not important. Finally, Jewish residents were more likely to use the clinic.

Comparisons With Other Evidence

Some additional data have come to my attention which permits further examination of the role of information. Efforts were made to encourage utilization of the newly established Bedford Health Center in the Bedford section of Brooklyn, New York. Some health aides were sent around to residences to inform persons of the availability of free services at the general medical clinic, at various dates during the period. Eligibility for the services was restricted to those residing in four health areas, and within those areas, aides were instructed to "knock on every door." The aides conducted a lengthy interview to determine needs for a broad range of social services and make necessary referrals, so that the information about the clinic would not have been foremost in the minds of many residents. This would presumably lead to lower estimates of the effects of health information than if counselling were more limited. Complete information on the impact of the counselling was not available, and only two of the four areas had enough data for adequate testing.[7]

Table 4 shows the mean number of admissions per week for the three health areas in which there was counselling, before, during and after the period in which counselling took place. In only one case was data for the period after counselling available and even then it only covered two weeks. To allow for a lag between counselling and admission the mean is also computed for the period during which counselling took place, excluding the first two weeks. In health areas 20 and 21 combined, admissions per week were 1/3 higher in the period after the first two weeks of interviewing than before interviewing began. This counselling would, therefore, increase the total patient load at the clinic by about 2 1/2 percent, assuming that the admission rate fell back to earlier levels after interviewing terminated. The counselling covered an area with about 40 percent of the households in the four health areas.

I know of few studies for which suitable comparisons with the present research can be made. Two recent studies of younger populations in different institutional settings are of interest. One is the portion of the Yale Ambulatory Care Studies which pertains to emergency room use which was recently published by Weinerman, et. al. The other is a just published study of utilization of prenatal clinics and the Judson Health Center in New York City conducted by Morton Silver as part of the Gouverneur Economic Research Project.[8] Both studies were analytical rather than descriptive, formulating hypotheses in advance, and subjecting them to multivariate analysis.

The Weinerman group examined variation in the percentage of emergency room visits in the Yale-New Haven Medical Center which were classified as non-urgent by physicians. This can be taken as a measure of the demand for clinic services. Over 2,000 visits were analyzed. Because users of a specific institution rather than complete population were studied there is uncertainty as to what extent non-users receive better care or no care at all. Age, a measure of health status, was found to be quite important, as were measures of health status in the present study. In the Yale study it reflected the greater need of the youngest and oldest groups for true emergency care. Negroes were found to have a greater percent of visits non-urgent, perhaps because of poorer health. While lower income groups tended to have a higher percent non-urgent, the difference was only weakly statistically significant. However, no result for income is given with education held constant. It may be for this reason that educational differences are not important. A number of information variables were found to be significant. Persons with a short length of residence at their current address, self-referred persons, and persons without a regular physician were more likely to use the emergency rooms for non-urgent problems. There were no significant effects of sex or religion. Zbrowski found that "the Jew tends to manifest a future oriented anxiety as to the symptomatic and general meaning of pain experience."[9] Therefore it is not surprising that Jewish persons were more likely to use a health maintenance clinic than emergency room, relative to persons of other faiths.

Morton Silver investigated the determinants of number of prenatal care visits and month of first visit for a sample of 142 women using seven prenatal clinics in low income areas of New York City. Regressions with number of first visits as dependent variables were run with and without holding constant month of first visit. Two variables which can be run interpreted as measures of (actual or expected) health status -- gestation period and unsuccessful outcome of prior pregnancies, were very important. An income measure was not available and education, which was associated with fewer visits and a later month of first visit may reflect income. Education's association with fewer visits probably reflects use of other sources. When the dependent variable was month of first visit anywhere, education was positive but not significantly related to month of first visit. Whether or not a woman had children, a "price" variable reflecting the difficulty of making a visit, adversely affected care. Travel time had the expected sign but was only significant in separate regressions for women without children. Whether or not a woman was working, in part interpreted as reflecting information and awareness about the importance of care, was associated with greater demand for prenatal care.

The Silver study provides additional information on amount of use which merits further attention. It is possible to crudely estimate the relative importance of medical care at early stages and frequent care, given the stage of the condition in accounting for the effects of each variable.[10] The main effects of education (or income) and employment status are through the frequency of visit, while the outcome of prior pregnancies and the presence of children operate relatively more through month of first visit.

Silver also examined visits of 125 children age one to five to the Judson Health Center on the lower east side of New York City. Once again an income variable was not available and the relationship of parents' education to the number of visits per child was negative. The relationships of education to the number of ill child visits and the number of illnesses were also negative, but against well child visits education was not significant. Silver notes that the classification of visits was not clear cut, depending, for example, on whether the illness was discovered during a check up. This raises the possibility that the more educated mothers more often recognized illnesses or communicated information about them. This would result in an understatement of the effect of education on ill child visits. Length of residence again was important. The foreign born tended to have significantly fewer well child visits. The new residents of the area tended to have more ill child visits.

The additional information cited, either directly or with some reinterpretation, can be considered to be very consistent with the Queensbridge data.

The foregoing analysis should be of value in determining the kinds of data which are required for planning community health services, and in making gross estimates of utilization and population coverage. For more complete planning, however, much more detailed information on the demand for different types of services is required. Furthermore, in order to actively influence the completeness with which medical care is distributed, we will have to ascertain what is behind the broad health, information, price and other variables so that specific programs can be formulated. This challenge to research is, thus far, being met with only an embryonic response.

FOOTNOTES

1. Nicetas H. Kuo *et. al.*, "The Queensbridge Health Maintenance Service for the Elderly, An Evaluation", unpublished manuscript, New York City Department of Health, n.d.

2. The President's Commission on Heart, Cancer and Stroke, *A National Program to Conquer Heart Disease, Cancer and Stroke*, Washington: 1964.

3. See J. Johnston, *Econometric Methods*, New York McGraw-Hill Book Company, Inc., 1963, pp. 221-228 and Arthur S. Goldberger, *Topics in Regression Analysis*, New York: The Macmillan Company, 1968, pp. 112-118.

4. Daniel B. Suits, "Use of Dummy Variables in Regression Equation", *Journal of the American Statistical Association, 52 (1957)*, pp. 548-51.

5. The standard test would be an F-test to see if the increase in explained variance when a set of dummy variables is added is significant in comparison with the unexplained variance. See Emanuel Melichar, "Least Squares Analysis of Economic Survey Data", *Proceedings of the Business and Economic Statistics Section of the American Statistical Association*, 1965, pp. 373-85.

6. *Op. cit.*

7. The imposition of fees after the period analyzed would have made interpretation of subsequent data impossible.

8. E. Richard Weinerman, Robert S. Ratner, A. Anthony Robbins, and Marvin A. Lavenhar, "Determinants of Use of Hospital Emergency Services", *American Journal of Public Health*, 56, No. 7 (July, 1966), pp. 1037-56, and Morton Silver, *Utilization of a Neighborhood Health Center: An Economic Model of Demand for Services Available at the Judson Health Center*, New York: Gouveneur Economic Research Project, 1968.

9. Mark Zborowski, "Cultural Components in Responses to Pain," *Journal of Social Issues*, VIII No. 4 (1952), p. 23.

10. Let the stage in the episode of which medical care is sought be determined by a linear function such as the following:

$$S = a_{10} + b_{11} X_1 + b_{12} X_2 + b_{13} X_3$$

The number of treatments depends on a number of factors including stage of first treatment.

$$N = a_{20} + b_{21} X_1 + b_{22} X_2 + b_{24} X_4 + b_s S$$

Solving,

$$N = a_{20} + b_s a_{10} + (b_{21} + b_s b_{11}) X_1$$

$$+ (b_{22} + b_s b_{12}) X_2 + b_s b_{13} X_3 + b_{24} X_4$$

The proportion of the effect of a variable operating through stage of episode can be estimated by terms such as $\frac{b_s b_{11}}{b_{21} + b_s b_{11}}$

Attitudes Toward Characteristics of Common Stocks

A Summary Presentation

Haskel Benishay, Northwestern University*

I. Introduction and Summary

This study examines empirically the determinants of the rate of return on common stocks for seven cross sections of 111 common stocks for the years 1958-1964 in a multiple regression analysis. Theoretically, the common stock rate of return under investigation is the ratio of the expected income associated with a particular equity to the market value of that equity. This rate is empirically represented by a ratio of a weighted average of past annual earnings to current market value of the equity.

The empirical rate of return is bypothesized to be a function of two groups of variables: a corrective group and an explanatory one. (1) The corrective variables are employed to attenuate errors involved in the measurement of the empirical representation of expected income in the numerator of the rate of return. (2) The explanatory variables are presumed to represent factors which exert a real influence on the relative desirability of stocks. They are employed to provide an explanation of the rate of return based on preferences and aversions of people in the market to various attributes of common stocks. The variables employed within the two groups are:

Corrective:
- -Trend (growth) in the market value of equity.
- -The pay-out ratio or the ratio of dividends to earnings.

Explanatory:
- -The stability of the income (earnings) stream.
- -The stability of equity value (price stability).
- -The size of the firm.
- -The debt-equity ratio.
- -The skewness of the distribution of equity values.
- -The relation between the market value of equity and a market stock index which may be thought of as the "conformity" of the market value of the equity of a firm to the index of the values of the equities of the market as a whole.

All but the last two variables were employed in an earlier empirical study of cross section data for the years 1954-1957 [1]. The earlier study and a later discussion associated with some of its controversial aspects [2,8] provided a benchmark from which this study was launched. A controversial result of the earlier study was a regression finding of a market preference for the variability of common stock equities. It was a result fortified by consistency in the various regressions rather than a high level of significance in each one regression separately. In this study one of the major interests was whether a preference for variability will persist in new data and with the inclusion in the regressions of additional relevant variables. Another major interest, whetted in recent years by work on portfolio selection [4,10,11,15] was the response of the market to the conformity (non-conformity) of equity market value to the market index.

To satisfy the first interest a measure of the skewness (third moment) of the equity distribution is added onto the explanatory variables. To satisfy the second an additional dimension is incorporated into the study via the inclusion of a conformity measure represented by the coefficient of the linear relation between the firm's equity market value and the value of Standard and Poor's 425 stock index.

---

*The author is a professor of managerial economics at Northwestern University. Len Wiltberger helped in the collection of data and William Melberg assisted in data processing. Responsibility for errors is, of course, mine.

The new and most noteworthy results in this study, as told by multiple regressions, are:

- -Equities of non-conforming stocks sell at a premium. Or, the market prefers stocks which do not conform to the movement of the index.
- -Equities whose market values are variable sell at a premium. Or, the market prefers variability to stability of the market values of common stock equities.
- -Equities whose market values are positively skewed sell at a premium. Or, the market prefers stocks whose market values are positively skewed.

The other results as told by our regressions are:

- -The equities of large firms sell at a premium in the market.
- -The equities of firms with larger debt-equity ratios sell at a discount.
- -The equities of firms whose earnings are more stable tend to sell at a premium (a mixed result).

The two corrective variables, aimed at attenuating the errors of measurement of earnings, perform as expected:

- -The past growth in equity is negatively associated with the rate of return.
- -The pay-out ratio is negatively associated with the rate of return.

II. Variables Employed

Why should the independent variables be expected to account for differences in rates of return on equity? Brief answers are provided below for each variable. Some additional discussion follows for equity variability, equity skewness and conformity to the movement of the market, taken as a group.

A. The "Correctors."

The Rate of Return: is empirically measured by a ratio of a weighted mean of the earnings of the company in the nine years preceding and including the cross-section year to the arithmetic mean of the high and low of the market values of company equity during the cross-section year. The weights employed for the nine observations of the weighted mean of earnings decline exponentially as the observations recede into the past away from the cross-section year. The weights for the i'th year back, where i=1 refers to the cross-section year, are:

$$\frac{(.8)^i}{\sum_{i=1}^{9} (.8)^i}$$

Thus for the cross-section year, the weight is

$$(.8) \Big/ \sum_{i=1}^{9} (.8)^i$$

for the farthest year away,

$$(.8)^9 \Big/ \sum_{i=1}^{9} (.8)^i$$

The weighted average of company earnings represents, in theory, expected earnings (income) of the company. Needless to say, the expected earnings sought and the weighted mean of earnings employed are not the same. It is mandatory to emphasize that this creates a large sized and fundamental empirical problem. If rates of return were measured without error then the differences in the rates would be attributable wholly to characteristics of the company of the type represented by the explanatory variables below. But they are not. And hence the need for variables whose function is to attenuate errors.

Growth in Equity: The inclusion among the independent variables of the past growth in the value of the equity is expected to provide correction for the divergence between the lagging empirical measure of expected income (earnings) and true expected income (earnings) which changes commensurately with growth in equity. While the empirical representation of earnings in the numerator of the rate of return may not reflect a change in expected income, the market value of the equity reflects such a change rather quickly and perhaps concomitantly. If the change in expected income is upwards, the measured rate will be smaller than the true one. In a similar fashion, when expected income declines, the measured rate will exceed the true rate. Since the past growth in equity reflects, in most cases, the growth in expected income, its inclusion in the regression may serve to provide a measure reflecting the speed with which expected income has changed and as a possible consequence a measure of the extent of the divergence between measured earnings which we employ and expected earnings which we wish to employ. Thus the larger is the rate of growth in equity, the larger will be the divergence between true and measured earnings, the smaller will be the empirically measured rate of return and the greater will be the negative correlation between equity growth and the empirically measured rate of return.

Growth in equity is included in the regression only due to its presumed capability to correct for erroneously measured earnings in the numerator of the rate of return, not for its alleged capability to predict either future growth in earnings or future growth in equity. This is an important distinction to bear in mind since the ability of past equity growth to predict future earnings or equity growth is itself a daring and perhaps invalid hypothesis.[1]

The *empirical* measure of equity growth is a ratio. Its numerator: the regression coefficient of the simple linear regression of the annual highs and lows of equity values, on time, for the nine years preceding and including the cross-section year. Its denominator: the arithmetic mean of the eighteen equity observations used to compute the numerator.

Pay-out ratio: The pay-out ratio, the ratio of dividends to earnings, is included for its presumed capacity to correct for an error involved in the representation of expected earnings by a weighted average of book earnings in the numerator of the rate of return. The rationale for the corrective function of the pay-out ratio lies essentially in its informational value. When included in the regression along with the growth in equity variable, as is done in this study, the pay-out ratio represents supplementary information on expected income. Of two companies with the same past growth rate, the one with the higher past pay-out ratio has actually been the more successful company, in the sense that it is the one with higher expected income per dollar of equity. Growth of the company with the higher pay-out ratio is in fact higher since it is accompanied by high dividends. Thus, as before, a negative association is expected between the rate of return and the pay-out ratio.[2]

The pay-out ratio is defined *empirically* as the weighted mean of nine ratios of dividends to earnings, in the cross-section year and the eight years preceding the cross-section year. The weights employed are the same as those described in the definition of the rate of return.

B. Risk Variables

Earnings-Time-Stability: People are said to prefer stability to variability of the earnings of their equities. Thus, stability of earnings and

[1] We assign equity growth only a corrective function not a predictive one. We feel that equity growth has no predictive power. Recent work on serial correlation of stock price changes suggest that growth in consecutive periods may be non-correlated. [4]

[2] For discussions of the determination of dividends see Lintner [9], Gordon [6,7], Walter [16] and Modigliani and Miller [14]. For a controversial discussion of the function of the pay-out ratio in regressions whose objective is to explain earnings-price ratios see exchange between this author and Gordon [1,2,8]. For an exposition of the view that the pay-out ratio may contain information about earnings see discussion by Modigliani and Miller [13].

the rate of return are expected to be negatively related. For a given level of the capital structure (a given debt-equity ratio) the larger is the variability of earnings, over time, the larger is the firm's probability of failure and the less attractive is its equity.

The empirical measure selected to represent the earnings-time-stability is a ratio whose numerator is the arithmetic mean of earnings in the nine years preceding and including the cross-section year and whose denominator is the standard deviation of the deviations from the line of the regression of the nine earnings observations on time. This empirical measure is essentially the reciprocal of the coefficient of variation of earnings after having taken account of the growth (trend) in earnings.

Equity-Time-Stability: Do participants in the stock market shun or prefer variability of equity values? Are investors more attracted by a higher than usual price, than repelled by an equally probable lower than usual price? Are investors as a group, and on balance, speculative? Or, does caution reign supreme? Since empirical results of some previous studies dealing with this question indicate that people preferred price variability to price stability [1,3], but as the bulk of the accepted opinion in the field is that caution prevails, our *a priori* hypothesis is a two tailed one, i.e., that either speculation, or caution may predominate.[3]

The empirical representation of equity-time-stability is a reciprocal of a ratio whose numerator is the standard deviation of the deviations from the line of regression which is run to obtain the numerator of equity-growth, and whose denominator is the arithmetic mean of the eighteen equity values used in the same regression.

Size: The larger the firm, the more liquid its shares and the more 'perfect' its market. Also, the larger the firm the more likely it is to be known to the general investing public (household word), the more its record is likely to be common knowledge and the smaller the amount of effort necessary to acquire information about it. On these grounds it is hypothesized that investors prefer large to small firms and consequently it is expected that the rate of return and size will be negatively correlated.[4]

The empirical representation of size is the sum of the weighted means of firm's equity and firm's long term debt, both in the nine years preceding and including the cross-section year. The weights employed in both are described in connection with the empirical definition of the rate of return.[5]

The Debt-Equity Ratio: The more debt there is in the capital structure, beyond the optimum, the higher the risk of default.[6] If the optimum debt-equity ratio is determined by the response of management to size and stability of earnings, and due to the fact that both earnings stability and size are also held constant in the regressions, the debt-equity ratio will come to represent deviations from the optimum and thus will be

[3]It is useful to note at this point that this cannot be considered separately and independently from the conformity variable.

[4]This negative relation is one widely accepted on theoretical grounds and never, to this writer's knowledge, contradicted by empirical findings. In my previous study, the size result was the most statistically significant one [1].

[5]In a previous study equity alone represented size, not equity plus debt [1]. But if the sum of equity and debt is the relevant measure of size, and equity alone represents it empirically, then, for a given value of equity, the debt-equity ratio, which is also included in the regressions, becomes a complementary measure of size rather than the capitalization measure which it is intended to be. For example, if two companies' total equity values are the same, then the company with the larger debt equity ratio is necessarily the company whose size, as measured by total assets, is larger. As a consequence of these considerations we represent size in this study by equity plus debt. For the sake of comparison with the previous study, regressions run with weighted mean of equity alone as the measure of size are presented in the appendix.

[6]For a view which discounts the possibility of such an optimum see the cost of capital discussion by Modigliani and Miller [12]. For evaluation of theoretical work in this area and empirical results of regressions relating capital structure to cost of capital see my earlier discussion [2, pp. 213-215].

positively correlated with the rate of return.[7]

The debt-equity ratio is represented empirically by the ratio of weighted mean of debt to the weighted mean of equity, where both means are based on nine years preceding and including the cross-section year. The weights are provided in the definition of the numerator of the rate of return.

Equity-Time Skewness: The distribution of equity values may be skewed to the right, symmetrical, or skewed to the left. We advance the hypothesis that, for the market as a whole and on balance, people prefer their equity values positively skewed and that they prefer more positive skewness to less. The larger is the third moment (skewness) of the time distribution of equities, the more attractive is the firm and the smaller is the equity rate of return.[8]

Empirically the equity-time skewness is defined as a ratio whose numerator is the cubed root of one ninth of the sum of cubed deviations from the regression run to obtain equity growth and whose denominator is the arithmetic mean of the eighteen equity figures used in the same regression. The division by mean equity is intended to deflate for differential size effects.

An alternative empirical definition of equity-time skewness, used experimentally, is also a ratio. Its numerator is the square of one ninth of the sum of cubed deviations from the regression run to obtain equity growth and whose denominator is the denominator of equity-time-stability raised to the sixth power. This is the ratio of the square of the third moment to the cube of the variance, which is a third moment deflated by the variance.[9]

Conformity of Equity to Market: Do investors prefer the market values of their stocks to move with or against the movement of the market as a whole? It is hypothesized here that investors prefer non-conforming stocks who move counter to the market to stocks who move marketwise. Consequently, it is expected that non-conforming equities will sell at a premium and that the extent of conformity and the rate of return on equity will be positively related.[10]

The empirical measure of the conformity of the equity to the market is a ratio. Its numerator is the regression coefficient of the simple linear regression of the annual highs and lows of equity values on Standard and Poor's 425 Stock Index, both sets of observations for the nine years preceding and including the cross-section year. Its denominator is the arithmetic mean equity of the same eighteen equity observations. Henceforth, we refer to this measure as Equity-Index Coefficient.

An alternative empirical measure for the conformity of the firm's equity to the equities of the market as a whole has been employed experimentally and is presented in the appendix. It is a ratio. Its numerator is the regression coefficient of the simple linear regression of: the eight first differences of the nine arithmetic means of the annual highs and lows of company equity on: the eight first differences of Standard and Poor's 425 Stock Index. Both sets of data relate to the nine years preceding and including the cross-section year. Its denominator is the arithmetic mean of the nine equity observations (absolute levels) used to obtain the eight first

[7]If the variables which determine optimum debt-equity ratio are not included in the regressions (or are not correctly measured) then one would expect a negative association between the rate of return and the ratio of debt to equity. In this case, high risk is associated with a low rate of return since only the safer companies can "afford" to have higher debt-equity ratios and the debt-equity ratio represents absolute levels, not deviations from an optimum.

[8]We are aware of the fact that at least in pure logic, the preference profile of an individual can be such that he will prefer negative to positive skewness. However, none of the individuals we have asked, in non-technical terms, whether he prefers negative skewness, and we have asked very many, has answered in the affirmative. This point is discussed again at the end of this section and in the section of results, in conjunction with the discussion of variability of equity and the conformity of equity to the market index.

[9]For details see discussion of the β_1 quantity in the chapter on moments in Yule and Kendall [17].

[10]For integration of this hypothesized preference with the hypothesized preference of equity variability and equity skewness see extended paper.

differences of equity. Henceforth, we refer to this measure as Equity-Index-FD Coefficient.

A Comment on Two Equity Value Hypotheses: Investors may prefer variability to stability for samll portions of their wealth but at the same time, may prefer stability to variability for larger portions of their total wealth.[11] Given this possibility, a preference for variability and non-conformity are not incompatible,[12] and it becomes evident that investors may very well prefer stocks which are both variable and non-conforming, thereby affording "limited gambling" for each stock but lower variability for the portfolio taken as a whole.

III. Empirical Findings

One hundred and eleven companies were studied in seven cross-sections in the years 1958-1964. For each company in a given cross-section year, most variables were computed on the basis of their values in the cross-section year and in the eight years preceding it. For exsample, variables in the 1958 cross-section are computed on the basis of observations which extend from 1958 back to 1950; variables in the 1959 cross-section on the basis of observations extending back to 1951.

The principal source of data was Moody's Handbook of Widely Held Common Stocks [18]. The companies chosen were industrial companies with comprehensive and complete data for the years 1950-1964. They had common stocks but no preferred stocks outstanding. The source for the annual figures of the Standard and Poor's 425 Stock Index was 1965 Statistical Supplement to the Survey of Current Business, where the value for 1941-43 equals 10.

The various hypotheses were evaluated in cross-sectional multiple linear regressions in which logarithmic values were used for the variables which are "logarithmable," i.e., which do not have zero or negative values which make logarithmic transformation impossible. These variables are: the rate of return, earnings-time stability, equity-time stability and size.

The empirical results are presented below in the "Main Regressions" table. The first and second columns of the table refer, respectively, to the year and the exponential weighting system employed in the weighted variables. The last column provides the square of the multiple correlation coefficient. The first line for each year provides the regression coefficient, the second the t ratio, the third the partial correlation coefficient. Tables in the appendix are presented in a similar form.

MAIN REGRESSIONS TABLE 1

To provide a quick impression of the predominant direction of relationships, the results are summarized in "SUMMARY OF RESULTS" table.

SUMMARY OF RESULTS TABLE 2

It becomes immediately evident that the empirical results have generally taken their anticipated routes. The one exception is the zig-zaggy path of the mixed results of the earnings stability regressions.

A few comments may be in order as regards two potentially controversial interrelated results: a market preference for equity variability combined with a market preference for equity non-conformity to market index. We are aware that many researchers may consider such a combination inconsistent. For example, Richard Bower adjudged similar results, which emerged in his own recent study [3], incompatible.[13] He accepted as perfectly valid the preference for non-conformity. But rejected the preference for equity variability as theoretically incompatible with non-conformity and wrong in its own right.[14] He also rejected it on the empirical grounds of equity skewness omission. He argued that the apparent preference for variability may be attributable to a combination of probable positive correlation between equity

[11] A rationale for this hypothesis, based on a Friedman-Savage view of the world [5] is developed in another paper.

[12] They are, of course, inconsistent if investors are assumed to have aversion to variability for all sizes of investment across the board. This is a common assumption as can be attested by current work [3,8,10,15].

[13] His data (different than mine) also revealed a preference for equity variability and non-conformity at the same time.

[14] Within a portfolio context, the preference for non-conformity can be a consequence of basic assumptions. For discussion see [3,10,11,15].

TABLE 2
SUMMARY OF RESULTS

	sign and its frequency	preference, aversion or correction	statistical significance
equity growth	-(6)	corrects	high
pay-out ratio	-(7)	corrects	high
log earnings-time stability	-(4)	prefer	low
log equity-time stability	+(6)	avert	medium
log size	-(7)	prefer	high
debt-equity ratio	+(7)	avert	medium
equity-time skewness	-(7)	prefer	high
equity-index coefficient	+(6)	avert	high

skewness and equity variability and a probable market preference for positive skewness. He suggested that since equity skewness was not included in his regressions the apparent preference for variability emerging in his empirical work might have reflected a real preference for the positive skewness excluded from his regressions.

In our regressions, however, equity skewness is included and it is thus held constant and prevented from interfering with the equity variability result. Therefore, the combination of preference for both equity non-conformity and equity variability, in the same market, and at the same time, can be said to emerge in our regressions <u>not due to the effect of an excluded equity skewness</u> measure.

We do not consider the two results necessarily incompatible. We think that it is not impossible that investors may prefer variability and stability at the same time, stability, for large portions of their wealth or its total, variability for smaller portions of it. But we are also well aware of the serious empirical difficulties inherent in the investigation of such far reaching theoretical hypotheses in multiple regressions. Hence, given potential uncertainties surrounding these findings and their controversial nature, we believe it best to leave the discussion of these results open ended and our own minds not made up.

Appendix A

Different Regression Types (Sets)

In addition to the regressions presented in the body of the paper to which we shall refer as set 1.8, 1 for first and .8 for the exponential weighting system employed, we have run some other regressions to compare the performance of empirical candidates competing to fill the place of some theoretical variables. Specifically, we have run three additional sets, each with its own particular purpose.

Set 2, the first of the comparative runs, is identical in every respect but one with Set 1. The difference is in that the empirical representation of equity-time skewness is measured essentially as the third moment deflated by equity variance (the alternative measure of skewness mentioned earlier), rather than essentially as a third moment deflated by the mean of equities (the primary measure). This affords a comparison between two methods of deflating skewness. In Set 1 deflation is accomplished through division by actual size as measured by arithmetic mean equities. In Set 2, division by a transform of variability, as measured by equity variance, constitutes deflation.[15] From the point of view of relative size of the multiple explanation (R^2) and the t ratio of skewness, the alternative measure employed in Set 2 is inferior.

Set 3 differs from Set 1 in that the actual values of the rate of return, earnings-time stability, equity-time stability and size (debt + equity) are employed in the regressions rather than the logarithms of the actual values. Also all observations of the size variable (equity plus debt) are measured

[15] See empirical definitions for precise procedures of deflation.

in billions of dollars. This provides a comparison of the performance of regressions employing logarithms and of ones which do not. It appears clearly that the non-logarithmic regressions perform more poorly.

Set 4 differs from Set 1 in that size is measured empirically as 'equity' alone not as 'debt plus equity.' This affords a comparison between two empirical representations of size. The difference between the two sets emerges not in the performance of size but rather in the performance of the debt-equity ratio.[16] In Set 1, where debt plus equity is employed, the performance of the debt-equity ratio is better.

Finally, we have provided some comparative experimentation within the first set by running three additional experimental regressions. In one, equity-index coefficient is replaced by the equity-index FD coefficient.[17] In another, the equity-index coefficient is omitted. In the third, both the equity-index coefficient as well as the skewness variable are omitted. The basic findings do not change as a result of these experiments.

Different Weighting Schemes

Weights are employed in the numerator of the rate of return, in the payout ratio, in size and in the debt-equity ratio. Since no a priori knowledge is available on the appropriateness of alternative weighting systems for a study like ours, we employed experimentally a number of weighting systems to gain additional knowledge.

The first, an exponential scheme where $P = .8$, has already been spelled out in the body of the paper in connection with the empirical definition of the numerator of the rate of return. We refer to this weight as 8 in the second column of the following sets of regression tables.

The second is also a set of exponentially declining weights, referred to in the second column of the tables below as 5, where $P = .5$. In this system of weights the weight for the i^{th} year back (where the cross-section year is 1^{st} year back) is:

$$\frac{(.5)^i}{\sum_{i=1}^{9} (.5)^i}$$

The third is also an exponential set similar to the second, referred to as 2 in the regression tables, where $P = .2$. In this system the weight for the i^{th} year back is:[18]

$$\frac{(.2)^i}{\sum_{i=1}^{9} (.2)^i}$$

The fourth weighting scheme, to which we refer in the regression tables as H, is computed as follows: $(9+1)/(2)(9) = 10/18$, is the weight for the first (cross-section) year and $1/(2)(9) = 1/18$ for each of the preceding eight years. Generally, $(n+1)/2n$ and $1/2n$ respectively, where n is the number of years. This amounts simply to computing the arithmetic mean of two quantities: (1) the value in the cross-section year (2) the arithmetic mean of values in the cross-section year and the eight years preceding.

The fifth weighting scheme, referred to in the regressions below as R, is not the same for all weighted variables. The weights applied to the different weighted variables are:

Year / Variable	1	2	3	4	5	6	7	8	9
rate of return	$\frac{10}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$
pay-out ratio	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0	0	0	0	0	0
equity	1	0	0	0	0	0	0	0	0
debt	1	0	0	0	0	0	0	0	0

The variables were processed with these R weights to provide as close a

[16] See previous discussion.

[17] See empirical definitions of the conformity variable.

[18] In the empirical work this system of weights made for negative rates of return in the year 1962. Hence logarithmic regressions were not run in this year for Set 1, Set 2 and Set 4.

comparison as is possible with earlier work which employed regressions with variables so weighted [1].

As may be seen from preceding paragraphs we have run twenty experimental sets of regressions, four types and five weighting systems within each type. In the various regression tables we designate the type of regression run, (1,2,3,4), as first digit from left, the system of weights employed as second digit or letter, (8,5,2,H,R) after a point. For example, first type (set) of regression, weight system 2 is designated as 1.2; second type of regression of weight system H is designated as 2.H.

The type of regressions are presented below in the order: Set 1, Set 2, Set 3, Set 4. Within each type (set) the systems of weights are arranged in the order: .8, .5, .2, H, R.

For the four types of regressions, it appears that: (a) Deflation of skewness by convential size, measured as the mean of equities, provides a better regression performance than deflation by variance. (b) Regressions in which logs are employed seem to perform substantially better. (c) And finally, the representation of size by the sum of debt and equity rather than by equity alone makes for better results for the debt-equity variable. As for weighting systems, it can be said unequivocally that the exponential system 8, where P = .8, performed best, the exponential system 2, where P = .2, performed worst. The other three, 5, H, R, performed about midway between best and worst.

Appendix B

One additional regression was run experimentally in which the variables are those in Table 1 above, plus a variable representing Kurtosis, whose definition follows.

Equity Time Kurtosis: The added variable, Kurtosis, is defined empirically as a ratio whose numerator is the fourth power root of one-ninth of the sum of deviations, raised to the fourth power, from the regression run to obtain equity growth and whose denominator is the arithmetic mean of the eighteen equity figures used in the same regressions. The division by mean equity is intended to deflate for differential size effects.

The table below entitled KURTOSIS EXPERIMENT provides the results of this experimental regression.

It appears that the relationship of the equity-time stability variable with the rate of return becomes inconsistent as a result of the inclusion of 'equity-time Kurtosis'. The performance of the Kurtosis variable itself seems to suggest, on balance, a weak preference for Kurtosis. Five out of the seven years emerge with negative signs for this variable.

REFERENCES

[1] Haskel Benishay, "Variability in Earnings-Price Ratios of Corporate Equities," American Economic Review, Vol. 51, March 1961, pp. 81-94.

[2] ______________, "Variability in Earnings-Price Ratios of Corporate Equities: Reply," American Economic Review, Vol. 52, March 1962, pp. 209-16.

[3] Richard S. Bower, "Risk and the Valuation of Common Stock," Econometric Society Meetings, December 1966.

[4] Eugene F. Fama, "The Behavior of Stock-Market Prices," Journal of Business, Vol. 38, No. 1, January 1965, pp. 34-105.

[5] Milton Friedman and L.J.Savage, "Expected-Utility Hypothesis and the Measurability of Utility," Journal of Political Economy, Vol. 60, December 1952, pp. 463-74.

[6] Myron J. Gordon, "Dividends, Earnings and Stock Prices," Review of Economics and Statistics, Vol. 41, May 1959, pp. 99-105.

[7] ______________, "The Savings Investment and Valuation of a Corporation," Review of Economics and Statistics, Vol. 44, February 1962, pp. 37-51.

[8] ______________, "Variability in Earnings Price Ratios: Comment," American Economic Review, Vol. 52, March 1962, pp. 203-09.

[9] John Lintner, "Distribution of Income of Corporations Among Dividends, Retained Earnings and Taxes," American Economic Review, Vol. 46, No. 2, May 1956, pp. 97-113.

[10] ____________, "Security Prices, Risk, and Maximal Gains From Diversification," Journal of Finance, Vol. 20, December 1965, pp. 587-615.

[11] Harry Markowitz, Portfolio Selection: Efficient Diversification of Investments, New York, Wiley, 1959.

[12] Franco Modigliani and Merton H. Miller, "The Cost of Capital, Corporation Finance and the Theory of Investment," American Economic Review, Vol. 48, June 1958, pp. 261-97.

[13] ____________, "The Cost of Capital Corporate Finance and the Theory of Investment: Reply," American Economic Review, Vol. 49, September 1959, pp. 655-69.

[14] ____________, "Dividend Policy, Growth and the Valuation of Shares," The Journal of Business, Vol. 34, No. 4, October 1961, pp. 411-53.

[15] William F. Sharpe, "Capital Asset Prices: A Theory of Market Equilibrium Under Conditions of Risk," Journal of Finance, Vol. 19, No. 3, September 1964, pp. 425-442.

[16] James E. Walter, "Dividends, Policies and Common Stock Prices," Journal of Finance, Vol. 11, No. 1, March 1956, pp. 29-42.

[17] G. Undy Yule and M.G. Kendall, An Introduction to the Theory of Statistics. Hafner Publishing Company, New York, 1950, p. 160.

[18] Moody's Handbook of Widely Held Common Stocks. New York, 1950-1964.

MAIN REGRESSIONS

TABLE 1

Regression Coefficients, T Ratios, and Partial Correlation Coefficieints, Respectively.

Year	Wt	Equity Growth	Payout Ratio	Log Earnin Time Stabil	Log Equity Time Stabil	Log Size	Debt Equity Ratio	Equity Time Skew	Equity Index Coeff.	R Square
1958	8	-10.39889	-0.00174	0.10408	-0.16434	-0.09433	0.04150	-0.24134	43.37730	0.61630
		-7.34204	-2.63153	1.72086	-1.52781	-4.49209	2.21273	-3.42328	6.69223	
		-0.58801	-0.25214	0.16797	-0.14957	-0.40640	0.21402	-0.32102	0.55237	
1959	8	-8.22376	-0.00188	-0.02974	0.13363	-0.06834	0.03007	-0.19723	37.58380	0.57264
		-3.81098	-2.66084	-0.57689	1.15048	-3.27511	1.73848	-2.99104	3.34552	
		-0.35304	-0.25477	-0.05703	0.11318	-0.30847	0.16964	-0.28397	0.31445	
1960	8	-10.48662	-0.00227	0.03424	0.02176	-0.08725	0.03150	-0.15045	49.16741	0.65436
		-7.05673	-3.33035	0.65589	0.19125	-4.05677	1.99896	-1.97172	6.15629	
		-0.57276	-0.31317	0.06481	0.01893	-0.37273	0.19416	-0.19161	0.52049	
1961	8	-7.68837	-0.00238	-0.07535	0.26376	-0.06863	0.03738	-0.12944	36.82053	0.59225
		-3.60211	-2.94258	-1.40325	2.14281	-2.85453	1.76802	-1.41940	2.99721	
		-0.33593	-0.27973	-0.13762	0.20755	-0.27199	0.17244	-0.13917	0.28450	
1962	8	-8.96324	-0.00071	-0.04527	0.40254	-0.07150	0.04172	-0.07863	39.30833	0.59171
		-4.62244	-2.39235	-0.88846	3.22473	-3.25902	1.83646	-1.10710	4.14371	
		-0.41617	-0.23050	-0.08763	0.30417	-0.30710	0.17890	-0.10897	0.37958	
1963	8	-1.69989	-0.00073	-0.01980	0.16019	-0.06893	0.04881	-0.15468	2.83478	0.51390
		-0.68860	-2.67716	-0.45461	1.38014	-3.31942	1.72963	-1.67714	0.28289	
		-0.06802	-0.25623	-0.04497	0.13540	-0.31224	0.16880	-0.16382	0.02800	
1964	8	1.17657	-0.00078	-0.03841	0.16854	-0.08725	0.07670	-0.19745	-10.24210	0.55627
		0.85151	-3.34000	-0.90656	1.48020	-4.59920	2.66397	-2.15359	-1.51635	
		0.08401	-0.31399	-0.08940	0.14501	-0.41444	0.25505	-0.20855	-0.14848	

KURTOSIS EXPERIMENT

Regression Coefficients, T Ratios, and Partial Correlation Coefficients, Respectively.

Equity Growth	Payout Ratio	Log Earnin Time Stabil	Log Equity Time Stabil	Log Size	Debt Equity Ratio	Equity Time Skew	Equity Index Coeff.	Equity Time Kurtosis	Year, Wt and R^2
-10.58976	-0.00176	0.10017	-0.09338	-0.09366	0.04090	-0.25894	44.35743	0.08108	1958 8
-6.96871	-2.63933	1.62294	-0.41272	-4.42338	2.16335	-3.00098	6.27851	0.35698	0.61678
-0.56982	-0.25401	0.15942	-0.04103	-0.40285	0.21044	-0.28612	0.52984	0.03550	
-7.93830	-0.00184	-0.01854	-0.06628	-0.07100	0.03262	-0.16466	35.99852	-0.20435	1959 8
-3.65854	-2.60024	-0.35361	-0.31311	-3.38579	1.87257	-2.29037	3.18388	-1.12894	0.57796
-0.34208	-0.25049	-0.03516	-0.03114	-0.31927	0.18317	-0.22220	0.30201	-0.11163	
-10.24121	-0.00224	0.03687	-0.09175	-0.08814	0.03237	-0.14118	47.87560	-0.13271	1960 8
-6.48566	-3.26108	0.69974	-0.34574	-4.06719	2.03242	-1.78570	5.65384	-0.47390	0.65512
-0.54224	-0.30865	0.06946	-0.03438	-0.37514	0.19822	-0.17494	0.49031	-0.04710	
-6.72279	-0.00216	-0.04987	-0.28774	-0.07038	0.04097	-0.07440	31.53142	-0.65241	1961 8
-3.12466	-2.69591	-0.91929	-0.98530	-2.97251	1.96230	-0.79494	2.55188	-2.07547	0.60893
-0.29690	-0.25909	-0.09109	-0.09757	-0.28363	0.19164	-0.07885	0.24611	-0.20225	
-8.54954	-0.00068	-0.04437	0.01630	-0.07166	0.04341	-0.08310	37.52073	-0.47286	1962 8
-4.35811	-2.28444	-0.87324	0.04899	-3.27546	1.91289	-1.17190	3.92180	-1.25167	0.59794
-0.39785	-0.22166	-0.08656	0.00487	-0.30988	0.18698	-0.11582	0.36353	-0.12359	
-1.68360	-0.00073	-0.02034	0.13095	-0.06884	0.04900	-0.15063	2.78033	-0.04097	1963 8
-0.67747	-2.65220	-0.46170	0.44878	-3.29656	1.72469	-1.50923	0.27577	-0.10934	0.51396
-0.06726	-0.25517	-0.04589	0.04461	-0.31168	0.16914	-0.14851	0.02743	-0.01088	
1.30824	-0.00079	-0.03915	0.22492	-0.08736	0.07662	-0.21151	-10.90567	0.07591	1964 8
0.88554	-3.32303	-0.91774	0.91841	-4.58256	2.64902	-1.98121	-1.50458	0.26038	0.55657
0.08777	-0.31394	-0.09094	0.09101	-0.41489	0.25488	-0.19342	-0.14806	0.02590	

REGRESSION COEFFICIENTS, T RATIOS, AND PARTIAL CORRELATION COEFFICIENTS, RESPECTIVELY.

YEAR WT	EQUITY GROWTH	PAYOUT RATIO	LOG EARNIN TIME STABIL	LOG EQUITY TIME STABIL	LOG SIZE	DEBT EQUITY RATIO	EQUITY TIME SKEW	EQUITY INDEX COEFF.	R SQUARE
SET 1.5	X2	X3	LOG X4	LOG X5	LDB+EQ	X7	X5 M3	X2SP	
1958 8	-10.39889	-0.00174	0.10408	-0.16434	-0.09433	0.04150	-0.24134	43.37730	0.61630
	-7.34204	-2.63153	1.72086	-1.52781	-4.49209	2.21273	-3.42328	6.69223	
	-0.58801	-0.25214	0.16797	-0.14957	-0.40640	0.21402	-0.32102	0.55237	
1959 8	-8.22376	-0.00188	-0.02974	0.13363	-0.06834	0.03007	-0.19723	37.58380	0.57264
	-3.81098	-2.66084	-0.57689	1.15048	-3.27511	1.73848	-2.99104	3.34552	
	-0.35304	-0.25477	-0.05703	0.11318	-0.30847	0.16964	-0.28397	0.31445	
1960 8	-10.48662	-0.00227	0.03424	0.02176	-0.08725	0.03150	-0.15045	49.16741	0.65436
	-7.05673	-3.33035	0.65589	0.19125	-4.05677	1.99896	-1.97172	6.15629	
	-0.57276	-0.31317	0.06481	0.01893	-0.37273	0.19416	-0.19161	0.52049	
1961 8	-7.68837	-0.00238	-0.07535	0.26376	-0.06863	0.03738	-0.12944	36.82053	0.59225
	-3.60211	-2.94258	-1.40325	2.14281	-2.85453	1.76802	-1.41940	2.99721	
	-0.33593	-0.27973	-0.13762	0.20755	-0.27199	0.17244	-0.13917	0.28450	
1962 8	-8.96324	-0.00071	-0.04527	0.40254	-0.07150	0.04172	-0.07863	39.30833	0.59171
	-4.62244	-2.39235	-0.88846	3.22473	-3.25902	1.83646	-1.10710	4.14371	
	-0.41617	-0.23050	-0.08763	0.30417	-0.30710	0.17890	-0.10897	0.37958	
1963 8	-1.69989	-0.00073	-0.01980	0.16019	-0.06893	0.04881	-0.15468	2.83478	0.51390
	-0.68860	-2.67716	-0.45461	1.38014	-3.31942	1.72963	-1.67714	0.28289	
	-0.06802	-0.25623	-0.04497	0.13540	-0.31224	0.16880	-0.16382	0.02800	
1964 8	1.17657	-0.00078	-0.03841	0.16854	-0.08725	0.07670	-0.19745	-10.24210	0.55627
	0.85151	-3.34000	-0.90656	1.48020	-4.59920	2.66397	-2.15359	-1.51635	
	0.08401	-0.31399	-0.08940	0.14501	-0.41444	0.25505	-0.20855	-0.14848	

YEAR WT	EQUITY GROWTH	PAYOUT RATIO	LOG EARNIN TIME STABIL	LOG EQUITY TIME STABIL	LOG SIZE	DEBT EQUITY RATIO	EQUITY TIME SKEW	EQUITY INDEX COEFF.	R SQUARE
SET 1.5	X2	X3	LOG X4	LOG X5	LDB+EQ	X7	X5 M3	X2SP	
1958 5	-7.74368	-0.00222	0.08324	-0.03833	-0.09505	0.02681	-0.16536	33.60322	0.44006
	-5.23404	-3.84782	1.32543	-0.34536	-4.37406	1.30473	-2.26521	4.96751	
	-0.46013	-0.35603	0.13012	-0.03418	-0.39742	0.12812	-0.21885	0.44136	
1959 5	-6.56094	-0.00213	-0.00911	0.17579	-0.07566	0.02879	-0.16226	31.47809	0.46816
	-3.24905	-3.60413	-0.18851	1.62207	-3.82841	1.90198	-2.62620	2.99316	
	-0.30625	-0.33610	-0.01866	0.15858	-0.35446	0.18507	-0.25166	0.28415	
1960 5	-8.46466	-0.00274	0.03689	0.09448	-0.07971	0.02732	-0.10481	40.84547	0.51718
	-5.58801	-4.79271	0.69283	0.81179	-3.58106	1.90131	-1.33309	5.02101	
	-0.48413	-0.42872	0.06844	0.08012	-0.33419	0.18501	-0.13086	0.44517	
1961 5	-6.57593	-0.00238	-0.04226	0.33823	-0.05410	0.04376	-0.08241	33.34312	0.41250
	-2.95839	-3.00242	-0.74349	2.60945	-2.15257	1.66094	-0.85810	2.60871	
	-0.28111	-0.28496	-0.07342	0.25016	-0.20845	0.16228	-0.08466	0.25009	
1962 5	-7.74582	-0.00042	0.01486	0.46738	-0.05291	0.07383	-0.05695	35.66093	0.35354
	-3.37912	-2.34398	0.24684	3.15026	-2.07855	2.05448	-0.67554	3.17669	
	-0.31729	-0.22608	0.02443	0.29777	-0.20158	0.19934	-0.06674	0.30005	
1963 5	0.76471	-0.00080	-0.04212	0.24668	-0.04623	0.11101	-0.14462	-4.98615	0.40292
	0.31301	-4.72346	-0.98229	2.15772	-2.31848	2.70795	-1.58051	-0.50274	
	0.03098	-0.42365	-0.09680	0.20893	-0.22374	0.25898	-0.15461	-0.04972	
1964 5	0.22888	-0.00114	-0.08985	0.19764	-0.07365	0.10325	-0.20037	-3.13857	0.51970
	0.16317	-6.79189	-2.11621	1.71796	-3.90234	2.81949	-2.15878	-0.45661	
	0.01615	-0.55805	-0.20508	0.16769	-0.36042	0.26889	-0.20903	-0.04516	

YEAR WT	EQUITY GROWTH	PAYOUT RATIO	LOG EARNIN TIME STABIL	LOG EQUITY TIME STABIL	LOG SIZE	DEBT EQUITY RATIO	EQUITY TIME SKEW	EQUITY INDEX COEFF.	R SQUARE
SET 1.2	X2	X3	LOG X4	LOG X5	LDB+EQ	X7	X5 M3	X2SP	
1958 2	-4.76930	-0.00267	0.08145	0.07413	-0.09270	0.01387	-0.10942	21.05067	0.32904
	-2.83966	-4.69133	1.13809	0.58780	-3.75521	0.51981	-1.31487	2.73761	
	-0.27067	-0.42128	0.11198	0.05810	-0.34851	0.05140	-0.12910	0.26162	
1959 2	-6.18963	-0.00284	0.01215	0.18632	-0.08126	0.02961	-0.14670	30.26309	0.47647
	-3.07965	-5.34268	0.25359	1.73216	-4.08690	2.17694	-2.39192	2.89001	
	-0.29167	-0.46761	0.02510	0.16904	-0.37511	0.21071	-0.23046	0.27511	
1960 2	-7.07367	-0.00371	0.03621	0.12455	-0.07127	0.02436	-0.07238	33.99409	0.49775
	-4.39048	-7.39963	0.63808	1.00058	-2.97523	1.76432	-0.85809	3.92620	
	-0.39868	-0.59102	0.06305	0.09859	-0.28258	0.17209	-0.08466	0.36234	
1961 2	-6.38104	-0.00220	0.00314	0.36322	-0.03555	0.05145	-0.05690	33.19214	0.28042
	-2.54004	-2.62189	0.04821	2.43846	-1.27398	1.23786	-0.51565	2.29967	
	-0.24391	-0.25128	0.00477	0.23470	-0.12515	0.12166	-0.05099	0.22202	
1963 2	2.06833	-0.00104	-0.08001	0.26418	-0.02751	0.13365	-0.13818	-9.68620	0.49123
	0.83054	-7.90411	-1.86820	2.27562	-1.38625	2.73239	-1.49009	-0.95779	
	0.08196	-0.61632	-0.18189	0.21981	-0.13598	0.26116	-0.14596	-0.09441	
1964 2	-0.66525	-0.00178	-0.12269	0.15395	-0.06981	0.08335	-0.23105	1.71667	0.66389
	-0.46512	-11.58707	-2.86915	1.31777	-3.65185	2.29377	-2.45684	0.24451	
	-0.04601	-0.75384	-0.27327	0.12938	-0.34004	0.22148	-0.23637	0.02420	
SET 1.H	X2	X3	LOG X4	LOG X5	LDB+EQ	X7	X5 M3	X2SP	
1958 H	-7.24257	-0.00200	0.10094	-0.10266	-0.09604	0.04049	-0.18497	29.32856	0.55667
	-5.28441	-3.50482	1.71418	-0.98404	-4.73411	1.82845	-2.70254	4.67297	
	-0.46361	-0.32785	0.16734	-0.09698	-0.42443	0.17815	-0.25850	0.41992	
1959 H	-7.87694	-0.00238	-0.00903	0.12635	-0.07504	0.03067	-0.16513	36.12269	0.60285
	-3.98172	-3.82126	-0.19169	1.19052	-3.86891	2.20384	-2.74352	3.50234	
	-0.36677	-0.35388	-0.01898	0.11707	-0.35773	0.21320	-0.26215	0.32764	
1960 H	-8.16756	-0.00265	0.03219	0.02400	-0.07349	0.02626	-0.11028	36.96612	0.57991
	-5.25892	-4.38798	0.58742	0.20065	-3.23323	1.76299	-1.36674	4.42410	
	-0.46185	-0.39849	0.05807	0.01986	-0.30489	0.17196	-0.13411	0.40124	
1961 H	-7.46470	-0.00222	-0.02492	0.26061	-0.05955	0.05380	-0.10596	36.23143	0.52117
	-3.40220	-2.69213	-0.44175	2.02078	-2.43212	1.66552	-1.11647	2.87328	
	-0.31924	-0.25757	-0.04370	0.19620	-0.23412	0.16271	-0.10988	0.27364	
1962 H	-8.35510	-0.00046	0.02081	0.37039	-0.06652	0.06210	-0.08995	36.43192	0.45494
	-3.52687	-2.75914	0.33348	2.41314	-2.51001	1.79298	-1.03382	3.14265	
	-0.32969	-0.26354	0.03300	0.23239	-0.24119	0.17480	-0.10183	0.29712	
1963 H	-0.76532	-0.00074	-0.05845	0.13490	-0.05505	0.06386	-0.15143	-0.53195	0.51866
	-0.32117	-4.10812	-1.40219	1.20112	-2.78372	1.88541	-1.69882	-0.05494	
	-0.03178	-0.37679	-0.13752	0.11810	-0.26572	0.18351	-0.16588	-0.00544	
1964 H	-0.81946	-0.00100	-0.08753	0.09712	-0.08869	0.06783	-0.23531	0.19442	0.58680
	-0.61076	-4.61556	-2.16157	0.88294	-4.85517	2.46916	-2.65883	0.02959	
	-0.06036	-0.41566	-0.20929	0.08709	-0.43327	0.23749	-0.25459	0.00293	

YEAR WT	EQUITY GROWTH	PAYOUT RATIO	LOG EARNIN TIME STABIL	LOG EQUITY TIME STABIL	LOG SIZE	DEBT EQUITY RATIO	EQUITY TIME SKEW	EQUITY INDEX COEFF.	R SQUARE
SET 1. R	X2	X3 AER	LOG X4	LOG X5	LDBEQA	X7 AER	X5 M3	X2SP	
1958 R	-7.51103	-0.00224	0.09782	-0.11468	-0.09382	0.03847	-0.16358	30.42139	0.56954
	-5.48105	-4.22515	1.69969	-1.12545	-4.69177	1.58833	-2.43602	4.87237	
	-0.47699	-0.38594	0.16596	-0.11075	-0.42131	0.15536	-0.23448	0.43451	
1959 R	-8.99776	-0.00207	-0.01986	0.10762	-0.07496	0.02856	-0.17516	41.94851	0.59893
	-4.52157	-3.78768	-0.41743	1.01512	-3.79773	2.34248	-2.88370	4.05339	
	-0.40862	-0.35115	-0.04130	0.10001	-0.35197	0.22594	-0.27456	0.37247	
1960 R	-8.70570	-0.00211	0.04213	0.03125	-0.07460	0.02461	-0.11052	40.65164	0.54838
	-5.41591	-3.37305	0.74631	0.25201	-3.13522	1.85693	-1.31697	4.71552	
	-0.47259	-0.31678	0.07370	0.02494	-0.29648	0.18083	-0.12930	0.42306	
1961 R	-7.54165	-0.00143	-0.01649	0.20743	-0.05611	0.03217	-0.11334	36.95877	0.49656
	-3.35320	-1.88073	-0.28881	1.57090	-2.27508	0.69504	-1.13057	2.85627	
	-0.31510	-0.18307	-0.02859	0.15369	-0.21976	0.06866	-0.11125	0.27214	
1962 R	-7.89774	-0.00076	0.01224	0.38990	-0.05938	0.08130	-0.07702	34.29721	0.45483
	-3.34639	-2.85169	0.19819	2.54332	-2.32745	1.77020	-0.88115	2.96832	
	-0.31453	-0.27173	0.01962	0.24420	-0.22457	0.17264	-0.08692	0.28198	
1963 R	-0.51548	-0.00049	-0.06123	0.19748	-0.05196	0.14682	-0.12906	-0.90066	0.48583
	-0.20906	-2.70318	-1.44559	1.72156	-2.64586	2.87698	-1.40726	-0.09006	
	-0.02070	-0.25855	-0.14169	0.16804	-0.25343	0.27396	-0.13801	-0.00892	
1964 R	-1.21774	-0.00046	-0.09306	0.14293	-0.08992	0.09070	-0.20019	2.89598	0.55264
	-0.86598	-3.37239	-2.20958	1.25959	-4.78827	2.72672	-2.19542	0.41992	
	-0.08543	-0.31673	-0.21373	0.12376	-0.42840	0.26065	-0.21242	0.04154	

SET 2. S	X2	X3	LOG X4	LOG X5	LDB+EQ	X7	X5M3V	X2SP	
1958 8	-9.48430	-0.00171	0.11077	-0.19413	-0.09269	0.04182	-0.04659	38.00611	0.59042
	-6.00591	-2.50469	1.72898	-1.55325	-4.26333	2.14813	-2.12907	5.16388	
	-0.51113	-0.24071	0.16874	-0.15201	-0.38890	0.20804	-0.20628	0.45524	
1959 8	-7.67798	-0.00180	-0.04898	0.20271	-0.06349	0.02764	-0.03802	34.70408	0.54449
	-3.44468	-2.47646	-0.92448	1.67123	-2.95801	1.54966	-1.44544	2.98764	
	-0.32281	-0.23815	-0.09116	0.16326	-0.28108	0.15166	-0.14168	0.28367	
1960 8	-10.30285	-0.00222	0.02125	0.10875	-0.08161	0.03107	-0.00526	48.17162	0.64124
	-6.81242	-3.19212	0.39316	0.95935	-3.74663	1.93507	-0.12175	5.92399	
	-0.55921	-0.30137	0.03890	0.09456	-0.34781	0.18818	-0.01205	0.50595	
1961 8	-7.16658	-0.00233	-0.07103	0.27394	-0.06929	0.03875	-0.03283	33.68741	0.58927
	-3.41578	-2.85570	-1.28173	2.21236	-2.87158	1.83053	-1.12248	2.79337	
	-0.32038	-0.27209	-0.12590	0.21398	-0.27349	0.17834	-0.11046	0.26658	
1962 8	-8.98176	-0.00067	-0.05541	0.43458	-0.06723	0.04011	-0.02071	39.48691	0.58723
	-4.48885	-2.24261	-1.10109	3.55430	-3.08314	1.74670	-0.32637	4.04128	
	-0.40615	-0.21677	-0.10838	0.33197	-0.29197	0.17042	-0.03230	0.37151	
1963 8	-1.79682	-0.00065	-0.03047	0.23131	-0.06208	0.03947	-0.03329	3.10565	0.50261
	-0.71422	-2.38507	-0.70023	2.14619	-3.01759	1.40304	-0.65746	0.30421	
	-0.07054	-0.22983	-0.06917	0.20786	-0.28628	0.13760	-0.06496	0.03011	
1964 8	1.45635	-0.00068	-0.05106	0.25345	-0.07939	0.06706	-0.03685	-11.85507	0.54370
	1.04389	-2.93478	-1.20515	2.43913	-4.24571	2.32498	-1.30384	-1.74456	
	0.10281	-0.27904	-0.11849	0.23476	-0.38754	0.22434	-0.12804	-0.17022	

YEAR WT	EQUITY GROWTH	PAYOUT RATIO	LOG EARNIN TIME STABIL	LOG EQUITY TIME STABIL	LOG SIZE	DEBT EQUITY RATIO	EQUITY TIME SKEW	EQUITY INDEX COEFF.	R SQUARE
SET 25	X2	X3	LOG X4	LOG X5	LDB+EQ	X7	X5M3V	X2SP	
1958 5	-7.49000	-0.00217	0.07860	-0.01746	-0.09277	0.02634	-0.01799	31.82956	0.41559
	-4.59815	-3.67897	1.19677	-0.13698	-4.17254	1.24988	-0.80405	4.19729	
	-0.41436	-0.34227	0.11767	-0.01356	-0.38184	0.12282	-0.07936	0.38377	
1959 5	-6.15982	-0.00203	-0.02679	0.24529	-0.07124	0.02755	-0.02449	29.37423	0.43770
	-2.96425	-3.35074	-0.54128	2.17408	-3.51868	1.77061	-0.99836	2.71159	
	-0.28163	-0.31489	-0.05352	0.21045	-0.32900	0.17268	-0.09837	0.25930	
1960 5	-8.34700	-0.00268	0.02630	0.16148	-0.07526	0.02730	0.00347	40.22288	0.50880
	-5.46837	-4.65570	0.48181	1.40963	-3.37433	1.88382	0.07930	4.90406	
	-0.47614	-0.41864	0.04765	0.13823	-0.31689	0.18336	0.00785	0.43680	
1961 5	-6.24065	-0.00235	-0.03880	0.34379	-0.05452	0.04550	-0.02163	31.32816	0.41116
	-2.86333	-2.95284	-0.66354	2.64488	-2.16752	1.73674	-0.70895	2.50241	
	-0.27276	-0.28063	-0.06556	0.25334	-0.20984	0.16947	-0.07002	0.24050	
1962 5	-7.53867	-0.00040	0.00955	0.48012	-0.04960	0.07176	-0.04576	34.71595	0.35300
	-3.20426	-2.24465	0.16107	3.31873	-1.96360	1.98023	-0.60944	3.01763	
	-0.30241	-0.21696	0.01595	0.31218	-0.19085	0.19241	-0.06023	0.28628	
1963 5	0.65651	-0.00075	-0.05230	0.30812	-0.04047	0.09730	-0.04153	-4.65955	0.39240
	0.26506	-4.47333	-1.22566	2.91383	-2.04992	2.38297	-0.83039	-0.46341	
	0.02624	-0.40498	-0.12047	0.27721	-0.19892	0.22964	-0.08194	-0.04584	
1964 5	0.38143	-0.00107	-0.09863	0.26312	-0.06720	0.09266	-0.05062	-4.14183	0.51307
	0.27112	-6.43597	-2.34166	2.53520	-3.62642	2.53546	-1.79107	-0.60303	
	0.02684	-0.53741	-0.22587	0.24347	-0.33794	0.24349	-0.17462	-0.05960	
SET 2[illegible]	X2	X3	LOG X4	LOG X5	LDB+EQ	X7	X5M3V	X2SP	
1958 2	-4.89772	-0.00260	0.07174	0.12002	-0.09061	0.01296	-0.00091	21.38941	0.31768
	-2.67925	-4.54805	0.97208	0.83995	-3.63650	0.48022	-0.03643	2.51065	
	-0.25642	-0.41061	0.09581	0.08288	-0.33878	0.04750	-0.00361	0.24125	
1959 2	-5.90427	-0.00276	-0.00785	0.26920	-0.07654	0.02877	-0.01242	28.78584	0.44851
	-2.85788	-5.05668	-0.16029	2.40999	-3.76703	2.06121	-0.51061	2.67181	
	-0.27228	-0.44770	-0.01587	0.23211	-0.34947	0.19997	-0.05049	0.25575	
1960 2	-7.02541	-0.00365	0.02436	0.18677	-0.06721	0.02436	0.01950	33.77806	0.49499
	-4.35132	-7.31280	0.42150	1.53562	-2.81869	1.75923	0.41855	3.89065	
	-0.39568	-0.58648	0.04170	0.15032	-0.26882	0.17161	0.04141	0.35948	
1961 2	-6.13719	-0.00218	0.00394	0.37249	-0.03585	0.05482	-0.01103	31.74029	0.27926
	-2.49417	-2.59175	0.05888	2.49646	-1.28380	1.34370	-0.31887	2.24713	
	-0.23976	-0.24857	0.00583	0.23996	-0.12610	0.13188	-0.03156	0.21719	
1963 2	1.90799	-0.00101	-0.08944	0.32364	-0.02258	0.11988	-0.04011	-9.13765	0.48332
	0.75817	-7.70210	-2.09985	3.02434	-1.14785	2.44769	-0.78987	-0.89400	
	0.07486	-0.60640	-0.20356	0.28687	-0.11293	0.23554	-0.07797	-0.08817	
1964 2	-0.49332	-0.00170	-0.13234	0.22624	-0.06280	0.07284	-0.06117	0.57867	0.65927
	-0.34435	-11.23704	-3.12359	2.15348	-3.34148	2.00520	-2.13856	0.08260	
	-0.03408	-0.74375	-0.29547	0.20854	-0.31411	0.19474	-0.20716	0.00818	

YEAR WT	EQUITY GROWTH	PAYOUT RATIO	LOG EARNIN TIME STABIL	LOG EQUITY TIME STABIL	LOG SIZE	DEBT EQUITY RATIO	EQUITY TIME SKEW	EQUITY INDEX COEFF.	R SQUARE
SET2.H	X2	X3	LOG X4	LOG X5	LDB+EQ	X7	X5M3V	X2SP	
1958 H	-6.67411	-0.00192	0.10355	-0.11096	-0.09465	0.04082	-0.03028	25.90851	0.53442
	-4.39359	-3.28529	1.67630	-0.92399	-4.54663	1.79236	-1.44190	3.65668	
	-0.39892	-0.30934	0.16374	-0.09111	-0.41050	0.17474	-0.14134	0.34044	
1959 H	-7.50249	-0.00233	-0.02989	0.20857	-0.07018	0.02928	-0.02013	34.17332	0.57645
	-3.66756	-3.61846	-0.61808	1.88363	-3.51809	2.03883	-0.83615	3.20091	
	-0.34133	-0.33729	-0.06108	0.18335	-0.32896	0.19788	-0.08251	0.30213	
1960 H	-8.07173	-0.00258	0.01961	0.10011	-0.06873	0.02623	0.00961	36.48781	0.57241
	-5.15355	-4.24927	0.34919	0.85011	-3.01489	1.74604	0.21364	4.32717	
	-0.45452	-0.38781	0.03455	0.08388	-0.28605	0.17036	0.02115	0.39383	
1961 H	-7.01822	-0.00219	-0.02379	0.27636	-0.06016	0.05753	-0.02108	33.56344	0.51762
	-3.25410	-2.62554	-0.40860	2.12580	-2.44888	1.78997	-0.69658	2.71103	
	-0.30668	-0.25160	-0.04042	0.20597	-0.23565	0.17451	-0.06881	0.25925	
1962 H	-7.96729	-0.00043	0.01241	0.38829	-0.06098	0.05764	-0.08060	34.66271	0.45500
	-3.27894	-2.59598	0.20230	2.59674	-2.31973	1.65192	-1.03941	2.91999	
	-0.30880	-0.24895	0.02003	0.24902	-0.22386	0.16142	-0.10238	0.27775	
1963 H	-0.91933	-0.00068	-0.06960	0.20727	-0.04876	0.05368	-0.03340	-0.00550	0.50730
	-0.37910	-3.79381	-1.67425	2.00486	-2.48441	1.58186	-0.68338	-0.00056	
	-0.03751	-0.35165	-0.16354	0.19471	-0.23887	0.15474	-0.06751	-0.00006	
1964 H	-0.52771	-0.00088	-0.10072	0.18896	-0.08029	0.05805	-0.05076	-1.50978	0.57269
	-0.38891	-4.04969	-2.48326	1.88700	-4.43650	2.09802	-1.86180	-0.22803	
	-0.03848	-0.37217	-0.23877	0.18366	-0.40219	0.20339	-0.18129	-0.02257	
SET 2.R	X2	X3 AER	LOG X4	LOG X5	LDBEQA	X7 AER	X5M3V	X2SP	
1958 R	-7.02160	-0.00224	0.09992	-0.12251	-0.09257	0.03956	-0.02745	27.42268	0.55234
	-4.65347	-4.14376	1.66579	-1.04966	-4.53756	1.59756	-1.33711	3.91385	
	-0.41848	-0.37959	0.16274	-0.10338	-0.40982	0.15624	-0.13125	0.36134	
1959 R	-8.58097	-0.00193	-0.04212	0.20269	-0.06944	0.02763	-0.01651	39.82646	0.56819
	-4.15503	-3.40896	-0.85600	1.81795	-3.40591	2.18517	-0.68071	3.70441	
	-0.38047	-0.31981	-0.08445	0.17716	-0.31955	0.21147	-0.06725	0.34436	
1960 R	-8.60887	-0.00204	0.02697	0.11477	-0.06917	0.02455	0.01800	40.16058	0.54137
	-5.31774	-3.24715	0.46614	0.93917	-2.90576	1.83836	0.38582	4.62313	
	-0.46590	-0.30608	0.04611	0.09259	-0.27650	0.17908	0.03817	0.41622	
1961 R	-7.06033	-0.00141	-0.01261	0.22130	-0.05595	0.04228	-0.02701	34.07884	0.49403
	-3.20456	-1.85060	-0.21492	1.68336	-2.26276	0.94269	-0.87260	2.69107	
	-0.30244	-0.18024	-0.02128	0.16441	-0.21863	0.09294	-0.08608	0.25747	
1962 R	-7.51640	-0.00072	0.00687	0.40131	-0.05507	0.07870	-0.07982	32.54137	0.45633
	-3.11668	-2.71779	0.11288	2.69182	-2.17649	1.70802	-1.02977	2.76041	
	-0.29488	-0.25986	0.01118	0.25754	-0.21067	0.16675	-0.10144	0.26365	
1963 R	-0.66132	-0.00045	-0.06945	0.25093	-0.04719	0.13184	-0.03547	-0.43340	0.47839
	-0.26538	-2.48495	-1.64660	2.35101	-2.42852	2.58888	-0.70378	-0.04287	
	-0.02627	-0.23892	-0.16091	0.22672	-0.23380	0.24831	-0.06952	-0.00424	
1964 R	-1.01206	-0.00041	-0.10071	0.20146	-0.08332	0.08091	-0.05397	1.57569	0.54795
	-0.72274	-3.07083	-2.41226	1.95716	-4.53327	2.44243	-1.92656	0.23062	
	-0.07138	-0.29091	-0.23231	0.19025	-0.40950	0.23506	-0.18738	0.02283	

YEAR WT	EQUITY GROWTH	PAYOUT RATIO	EARNIN TIME STABIL	EQUITY TIME STABIL	SIZE	DEBT EQUITY RATIO	EQUITY TIME SKEW	EQUITY INDEX COEFF.	R SQUARE
SET 3.8	X2	X3	X4	X5	DB+EQB	X7	X5 M3	X2SP	
1958 8	-1.28223	-0.00030	-0.00038	-0.00240	-0.00065	0.00545	-0.02752	5.14518	0.55020
	-5.84419	-2.62563	-0.64072	-1.59901	-1.10298	0.97522	-2.54585	5.11386	
	-0.50085	-0.25161	-0.06331	-0.15638	-0.10857	0.09611	-0.24443	0.45174	
1959 8	-0.81849	-0.00024	-0.00060	0.00052	-0.00005	0.00058	-0.01721	3.48796	0.52353
	-2.99570	-2.56975	-1.87405	0.37935	-0.09818	0.11292	-2.19170	2.44443	
	-0.28437	-0.24659	-0.18244	0.03753	-0.00972	0.01118	-0.21207	0.23524	
1960 8	-1.03085	-0.00031	-0.00040	-0.00008	-0.00004	0.00014	-0.00324	4.53950	0.55389
	-4.93910	-3.05662	-1.28941	-0.05150	-0.07936	0.02372	-0.30040	3.99878	
	-0.43932	-0.28967	-0.12664	-0.00510	-0.00786	0.00235	-0.02973	0.36813	
1961 8	-0.65376	-0.00027	-0.00062	0.00141	-0.00024	0.00334	-0.00722	2.85496	0.51823
	-2.60964	-2.73840	-2.35806	0.92259	-0.48331	0.60985	-0.63513	1.97323	
	-0.25018	-0.26169	-0.22737	0.09097	-0.04780	0.06027	-0.06276	0.19175	
1962 8	-1.12362	-0.00009	-0.00034	0.00442	-0.00099	0.01119	0.00048	4.85466	0.53814
	-4.42391	-2.30043	-1.41776	2.62513	-2.16837	2.06896	0.05071	3.88721	
	-0.40123	-0.22209	-0.13902	0.25157	-0.20992	0.20069	0.00502	0.35920	
1963 8	-0.16715	-0.00010	-0.00021	0.00177	-0.00126	0.01456	-0.01502	0.12146	0.47456
	-0.50375	-2.81192	-1.44245	1.22125	-2.86420	2.52828	-1.25538	0.08976	
	-0.04982	-0.26822	-0.14139	0.12005	-0.27284	0.24284	-0.12335	0.00889	
1964 8	0.34281	-0.00011	-0.00043	0.00245	-0.00159	0.02372	-0.02125	-2.27659	0.53702
	2.17324	-3.74454	-2.21878	1.75065	-4.52444	4.17125	-2.04357	-2.93955	
	0.21037	-0.34764	-0.21457	0.17079	-0.40884	0.38174	-0.19832	-0.27946	
SET 3.5	X2	X3	X4	X5	DB+EQB	X7	X5 M3	X2SP	
1958 5	-1.05524	-0.00043	-0.00047	-0.00064	0.00001	-0.00225	-0.01851	4.38045	0.35750
	-4.52694	-4.36476	-0.75839	-0.41122	0.02803	-0.36638	-1.64683	4.10722	
	-0.40902	-0.39671	-0.07488	-0.04068	0.00278	-0.03625	-0.16093	0.37671	
1959 5	-0.64239	-0.00030	-0.00046	0.00161	-0.00030	0.00366	-0.01638	2.88585	0.39091
	-2.36872	-3.69378	-1.44732	1.16741	-0.70772	0.70152	-2.07880	2.03624	
	-0.22834	-0.34349	-0.14186	0.11483	-0.06990	0.06929	-0.20161	0.19764	
1960 5	-0.88197	-0.00042	-0.00033	0.00126	-0.00022	0.00268	-0.00075	4.03025	0.43863
	-4.12124	-5.00009	-1.04633	0.79093	-0.45239	0.47837	-0.06669	3.47047	
	-0.37782	-0.44368	-0.10305	0.07807	-0.04475	0.04731	-0.00660	0.32498	
1961 5	-0.55703	-0.00034	-0.00052	0.00288	-0.00029	0.00503	-0.00155	2.61033	0.37719
	-2.22059	-3.72831	-1.97571	1.87216	-0.64865	1.00681	-0.13463	1.80609	
	-0.21474	-0.34631	-0.19199	0.18227	-0.06409	0.09920	-0.01333	0.17604	
1962 5	-0.77395	-0.00006	-0.00029	0.00600	-0.00098	0.01397	0.00287	3.48774	0.34975
	-2.77528	-2.79669	-1.11557	3.23631	-2.03742	2.50753	0.27506	2.54393	
	-0.26497	-0.26687	-0.10979	0.30516	-0.19775	0.24097	0.02722	0.24426	
1963 5	0.29876	-0.00010	-0.00027	0.00252	-0.00141	0.02238	-0.01827	-1.48373	0.40917
	0.94106	-4.50365	-1.99158	1.82627	-3.35347	3.86977	-1.60772	-1.14697	
	0.09278	-0.40727	-0.19347	0.17794	-0.31513	0.35780	-0.15721	-0.11284	
1964 5	0.07332	-0.00011	-0.00060	0.00232	-0.00152	0.02844	-0.02482	-0.67002	0.46183
	0.43279	-5.07869	-2.84392	1.52209	-4.48490	4.25063	-2.19578	-0.80621	
	0.04281	-0.44926	-0.27105	0.14903	-0.40585	0.38792	-0.21245	-0.07957	
SET 3.2	X2	X3	X4	X5	DB+EQB	X7	X5 M3	X2SP	
1958 2	-0.68180	-0.00049	-0.00043	0.00041	0.00018	-0.00510	-0.01343	2.76662	0.28442
	-2.69129	-5.37341	-0.63343	0.24342	0.31960	-0.76456	-1.10084	2.38210	
	-0.25749	-0.46970	-0.06260	0.02409	0.03163	-0.07549	-0.10836	0.22956	
1959 2	-0.57744	-0.00040	-0.00039	0.00213	-0.00055	0.00690	-0.01422	2.63902	0.39434
	-1.99043	-5.04975	-1.15054	1.42734	-1.26264	1.28347	-1.67165	1.73820	
	-0.19336	-0.44721	-0.11319	0.13994	-0.12405	0.12607	-0.16330	0.16961	
1960 2	-0.72729	-0.00049	-0.00032	0.00145	-0.00029	0.00343	0.00197	3.25590	0.44638
	-3.23615	-6.94608	-0.96991	0.86552	-0.56700	0.63307	0.16665	2.66867	
	-0.30514	-0.56668	-0.09560	0.08539	-0.05605	0.06256	0.01650	0.25547	
1961 2	-0.57713	-0.00035	-0.00039	0.00325	-0.00019	0.00458	-0.00018	2.84582	0.30163
	-2.18040	-3.93630	-1.39838	1.98665	-0.38802	0.88512	-0.01503	1.86788	
	-0.21103	-0.36314	-0.13715	0.19301	-0.03839	0.08730	-0.00149	0.18186	
1962 2	-0.53092	-0.00006	-0.00025	0.00637	-0.00084	0.01700	0.00250	2.37823	0.26808
	-1.59173	-3.32148	-0.81299	2.87943	-1.31639	2.51803	0.20061	1.45082	
	-0.15568	-0.31241	-0.08024	0.27418	-0.12925	0.24192	0.01986	0.14219	
1963 2	0.41424	-0.00010	-0.00033	0.00195	-0.00118	0.02572	-0.01852	-1.89689	0.38290
	1.16398	-5.07237	-2.13175	1.26031	-2.41785	3.68480	-1.46370	-1.30803	
	0.11449	-0.44881	-0.20652	0.12383	-0.23282	0.34275	-0.14343	-0.12844	
1964 2	-0.23360	-0.00013	-0.00076	0.00116	-0.00153	0.03141	-0.03026	0.89083	0.47863
	-1.20723	-5.46373	-3.14088	0.65927	-4.22306	3.89161	-2.32704	0.93808	
	-0.11869	-0.47582	-0.29696	0.06514	-0.38578	0.35956	-0.22453	0.09249	

YEAR WT	EQUITY GROWTH	PAYOUT RATIO	EARNIN TIME STABIL	EQUITY TIME STABIL	SIZE	DEBT EQUITY RATIO	EQUITY TIME SKEW	EQUITY INDEX COEFF.	R SQUARE
SET 3.H	X2	X3	X4	X5	DB+EQB	X7	X5 M3	X2SP	
1958 H	-0.82625	-0.00035	-0.00029	-0.00136	-0.00058	0.00430	-0.02074	3.11105	0.47517
	-3.91408	-3.70925	-0.51076	-0.94450	-1.02149	0.77680	-2.01001	3.21098	
	-0.36136	-0.34475	-0.05051	-0.09311	-0.10063	0.07669	-0.19519	0.30299	
1959 H	-0.75972	-0.00031	-0.00052	0.00089	-0.00060	0.00659	-0.01396	3.22114	0.56418
	-2.95597	-3.62312	-1.71828	0.67342	-1.26939	1.31384	-1.86986	2.39300	
	-0.28090	-0.33767	-0.16773	0.06653	-0.12471	0.12900	-0.18205	0.23056	
1960 H	-0.81519	-0.00039	-0.00037	0.00014	-0.00018	0.00188	0.00095	3.40041	0.52272
	-3.90712	-4.63392	-1.20272	0.09135	-0.32323	0.33738	0.08761	2.99482	
	-0.36080	-0.41703	-0.11825	0.00904	-0.03199	0.03339	0.00867	0.28430	
1961 H	-0.64330	-0.00030	-0.00045	0.00161	-0.00020	0.00392	-0.00588	2.87552	0.47603
	-2.60283	-3.13371	-1.73748	1.05326	-0.40738	0.73346	-0.52205	2.01848	
	-0.24956	-0.29635	-0.16955	0.10373	-0.04030	0.07243	-0.05162	0.19598	
1962 H	-0.88445	-0.00006	-0.00029	0.00466	-0.00108	0.01436	-0.00024	3.74095	0.47421
	-3.17235	-3.16500	-1.11749	2.51933	-2.14925	2.36392	-0.02288	2.72924	
	-0.29967	-0.29904	-0.10998	0.24203	-0.20815	0.22790	-0.00227	0.26088	
1963 H	-0.08233	-0.00010	-0.00025	0.00088	-0.00137	0.02065	-0.01621	-0.20224	0.49606
	-0.25049	-3.95101	-1.75084	0.61578	-3.21853	3.26450	-1.38127	-0.15083	
	-0.02479	-0.36432	-0.17081	0.06086	-0.30364	0.30757	-0.13550	-0.01493	
1964 H	-0.00760	-0.00013	-0.00060	0.00105	-0.00169	0.02887	-0.02781	-0.54646	0.54208
	-0.04496	-4.28391	-2.84477	0.68846	-4.69449	4.27397	-2.45754	-0.65816	
	-0.00445	-0.39049	-0.27112	0.06801	-0.42151	0.38973	-0.23643	-0.06503	

YEAR WT	EQUITY GROWTH	PAYOUT RATIO	EARNIN TIME STABIL	EQUITY TIME STABIL	SIZE	DEBT EQUITY RATIO	EQUITY TIME SKEW	EQUITY INDEX COEFF.	R SQUARE
SET 3.R	X2	X3 AER	X4	X5	DBEQAB	X7 AER	X5 M3	X2SP	
1958 R	-0.92885	-0.00040	-0.00022	-0.00177	-0.00044	0.00333	-0.01740	3.50831	0.50217
	-4.41356	-4.70940	-0.39536	-1.27002	-0.95686	0.60056	-1.74046	3.65543	
	-0.40044	-0.42261	-0.03912	-0.12477	-0.09432	0.05936	-0.16983	0.34034	
1959 R	-0.92400	-0.00027	-0.00058	0.00073	-0.00051	0.00668	-0.01432	4.05316	0.56025
	-3.54264	-3.72547	-1.89870	0.54622	-1.33986	1.43600	-1.88037	2.98102	
	-0.33100	-0.34608	-0.18476	0.05401	-0.13151	0.14077	-0.18304	0.28309	
1960 R	-0.90291	-0.00030	-0.00036	0.00033	-0.00008	0.00119	0.00112	3.95983	0.47930
	-4.10512	-3.46095	-1.11697	0.20213	-0.15444	0.23633	0.09764	3.32653	
	-0.37655	-0.32418	-0.10993	0.02001	-0.01529	0.02339	0.00967	0.31284	
1961 R	-0.65084	-0.00019	-0.00041	0.00084	-0.00037	0.00129	-0.00671	2.94818	0.44844
	-2.56082	-2.18320	-1.54854	0.54485	-0.61142	0.24494	-0.57714	2.01133	
	-0.24578	-0.21129	-0.15156	0.05387	-0.06043	0.02425	-0.05705	0.19532	
1962 R	-0.84956	-0.00010	-0.00033	0.00454	-0.00086	0.01232	-0.00040	3.56071	0.47666
	-3.04540	-3.24820	-1.26455	2.46721	-1.59323	2.18651	-0.03856	2.59816	
	-0.28870	-0.30617	-0.12424	0.23731	-0.15583	0.21160	-0.00382	0.24914	
1963 R	-0.10119	-0.00007	-0.00028	0.00149	-0.00121	0.02511	-0.01295	-0.04975	0.45527
	-0.29554	-2.67268	-1.88361	1.00614	-2.58272	3.52415	-1.06320	-0.03569	
	-0.02925	-0.25583	-0.18334	0.09913	-0.24775	0.32946	-0.10469	-0.00353	
1964 R	-0.05641	-0.00006	-0.00067	0.00151	-0.00156	0.03342	-0.02418	-0.21667	0.51645
	-0.32700	-3.44842	-3.07635	0.97011	-4.96834	4.57456	-2.09276	-0.25630	
	-0.03236	-0.32313	-0.29139	0.09561	-0.44142	0.41260	-0.20290	-0.02537	

YEAR WT	EQUITY GROWTH	PAYOUT RATIO	LOG EARNIN TIME STABIL	LOG EQUITY TIME STABIL	LOG SIZE	DEBT EQUITY RATIO	EQUITY TIME SKEW	EQUITY INDEX COEFF.	R SQUARE
SET 4.8	X2	X3	LOG X4	LOG X5	LOG X6	X7	X5 M3	X2SP	
1958 8	-10.32769	-0.00168	0.10357	-0.16181	-0.09366	0.02564	-0.24302	43.11153	0.61683
	-7.30446	-2.52895	1.71535	-1.50541	-4.51086	1.45992	-3.44777	6.66338	
	-0.58604	-0.24290	0.16745	-0.14743	-0.40781	0.14307	-0.32307	0.55071	
1959 8	-8.12950	-0.00182	-0.03035	0.13608	-0.06668	0.01940	-0.19497	37.13144	0.57158
	-3.76454	-2.57580	-0.58815	1.16986	-3.23250	1.19399	-2.95628	3.30318	
	-0.34927	-0.24713	-0.05814	0.11506	-0.30483	0.11740	-0.28093	0.31086	
1960 8	-10.45802	-0.00220	0.03474	0.02634	-0.08468	0.01890	-0.14470	49.05161	0.65282
	-7.02456	-3.21771	0.66249	0.23107	-3.99148	1.28171	-1.89624	6.12891	
	-0.57100	-0.30357	0.06546	0.02287	-0.36755	0.12590	-0.18453	0.51880	
1961 8	-7.65023	-0.00234	-0.07458	0.26485	-0.06712	0.02609	-0.12651	36.61382	0.59176
	-3.58559	-2.88812	-1.38404	2.15030	-2.83152	1.31303	-1.38612	2.98129	
	-0.33457	-0.27495	-0.13577	0.20824	-0.26995	0.12892	-0.13597	0.28311	
1962 8	-8.93720	-0.00072	-0.04289	0.40327	-0.07193	0.02884	-0.07794	39.17143	0.59342
	-4.62268	-2.41558	-0.84146	3.23830	-3.33112	1.35356	-1.10081	4.14163	
	-0.41619	-0.23262	-0.08303	0.30533	-0.31323	0.13283	-0.10835	0.37942	
1963 8	-1.57853	-0.00073	-0.01772	0.15969	-0.07028	0.03453	-0.15664	2.34039	0.51719
	-0.64108	-2.69263	-0.40764	1.38105	-3.43329	1.29524	-1.70412	0.23416	
	-0.06335	-0.25761	-0.04033	0.13548	-0.32186	0.12721	-0.16638	0.02318	
1964 8	1.21060	-0.00078	-0.03625	0.16917	-0.08844	0.05775	-0.19897	-10.41056	0.56112
	0.88239	-3.34963	-0.86028	1.49540	-4.74495	2.11198	-2.18376	-1.55260	
	0.08704	-0.31480	-0.08487	0.14647	-0.42523	0.20469	-0.21134	-0.15195	
SET 4.5	X2	X3	LOG X4	LOG X5	LOG X6	X7	X5 M3	X2SP	
1958 5	-7.65368	-0.00218	0.08171	-0.03502	-0.09294	0.00958	-0.16555	33.23851	0.43790
	-5.16695	-3.75643	1.29953	-0.31491	-4.32049	0.49470	-2.26317	4.90901	
	-0.45546	-0.34861	0.12762	-0.03117	-0.39331	0.04892	-0.21866	0.43716	
1959 5	-6.45582	-0.00207	-0.00921	0.17949	-0.07410	0.01758	-0.15906	30.97986	0.46711
	-3.19504	-3.49498	-0.19040	1.65404	-3.79812	1.24023	-2.57541	2.94449	
	-0.30162	-0.32703	-0.01885	0.16162	-0.35200	0.12189	-0.24710	0.27990	
1960 5	-8.42515	-0.00268	0.03756	0.10022	-0.07722	0.01668	-0.09831	40.66815	0.51556
	-5.55746	-4.69177	0.70241	0.86028	-3.52714	1.24712	-1.25192	4.99424	
	-0.48210	-0.42131	0.06938	0.08487	-0.32971	0.12255	-0.12302	0.44327	
1961 5	-6.54888	-0.00235	-0.04146	0.33811	-0.05312	0.03340	-0.08089	33.19104	0.41230
	-2.94841	-2.96967	-0.72722	2.60806	-2.14385	1.33798	-0.84193	2.59877	
	-0.28024	-0.28210	-0.07182	0.25003	-0.20765	0.13133	-0.08308	0.24920	
1962 5	-7.70980	-0.00042	0.01541	0.46713	-0.05277	0.06133	-0.05701	35.48190	0.35392
	-3.36704	-2.35509	0.25585	3.14946	-2.09334	1.78117	-0.67648	3.16436	
	-0.31627	-0.22710	0.02532	0.29770	-0.20296	0.17368	-0.06683	0.29899	
1963 5	0.81758	-0.00080	-0.04165	0.24679	-0.04707	0.09780	-0.14518	-5.19434	0.40446
	0.33485	-4.72854	-0.97351	2.16219	-2.37779	2.46585	-1.58930	-0.52407	
	0.03314	-0.42402	-0.09595	0.20935	-0.22917	0.23719	-0.15545	-0.05182	
1964 5	0.23895	-0.00114	-0.08858	0.19865	-0.07463	0.08452	-0.20119	-3.17729	0.52313
	0.17118	-6.80104	-2.09574	1.73436	-4.00868	2.40412	-2.17722	-0.46457	
	0.01695	-0.55856	-0.20318	0.16925	-0.36892	0.23157	-0.21074	-0.04595	
SET 4.2	X2	X3	LOG X4	LOG X5	LOG X6	X7	X5 M3	X2SP	
1958 2	-4.68570	-0.00263	0.07942	0.07773	-0.09034	-0.00512	-0.10889	20.70403	0.32647
	-2.78514	-4.61975	1.10868	0.61513	-3.69560	-0.20254	-1.30601	2.68893	
	-0.26585	-0.41597	0.10912	0.06079	-0.34364	-0.02005	-0.12825	0.25728	
1959 2	-6.08536	-0.00278	0.01233	0.19134	-0.08004	0.01851	-0.14281	29.77537	0.47636
	-3.02839	-5.22854	0.25720	1.77790	-4.08393	1.46191	-2.33208	2.84454	
	-0.28722	-0.45975	0.02546	0.17337	-0.37488	0.14326	-0.22499	0.27110	
1960 2	-7.03078	-0.00367	0.03663	0.13066	-0.06875	0.01551	-0.06571	33.78861	0.49631
	-4.36169	-7.32578	0.64285	1.04899	-2.92158	1.21145	-0.78058	3.90021	
	-0.39648	-0.58716	0.06352	0.10331	-0.27789	0.11910	-0.07706	0.36025	
1961 2	-6.37679	-0.00219	0.00384	0.36205	-0.03550	0.04266	-0.05714	33.16552	0.28055
	-2.53967	-2.61349	0.05879	2.43141	-1.28155	1.05868	-0.51790	2.29916	
	-0.24387	-0.25052	0.00582	0.23406	-0.12588	0.10425	-0.05121	0.22197	
1963 2	2.08736	-0.00104	-0.07982	0.26434	-0.02808	0.12482	-0.13851	-9.75814	0.49167
	0.83827	-7.90549	-1.86654	2.27856	-1.41838	2.59904	-1.49470	-0.96504	
	0.08272	-0.61638	-0.18174	0.22008	-0.13908	0.24922	-0.14640	-0.09512	
1964 2	-0.65984	-0.00178	-0.12161	0.15529	-0.07073	0.06623	-0.23164	1.70394	0.66596
	-0.46350	-11.59805	-2.85585	1.33460	-3.74880	1.90275	-2.47303	0.24387	
	-0.04584	-0.75415	-0.27210	0.13101	-0.34799	0.18514	-0.23784	0.02414	

YEAR WT	EQUITY GROWTH	PAYOUT RATIO	LOG EARNIN TIME STABIL	LOG EQUITY TIME STABIL	LOG SIZE	DEBT EQUITY RATIO	EQUITY TIME SKEW	EQUITY INDEX COEFF.	R SQUARE
SET 4.H	X2	X3	LOG X4	LOG X5	LOG X6	X7	X5 M3	X2SP	
1958 H	-7.16931	-0.00196	0.10020	-0.09952	-0.09543	0.02142	-0.18532	29.04114	0.55735
	-5.23721	-3.42160	1.70505	-0.95476	-4.75404	1.02402	-2.70947	4.63370	
	-0.46035	-0.32087	0.16647	-0.09412	-0.42589	0.10088	-0.25912	0.41701	
1959 H	-7.78685	-0.00231	-0.00899	0.13010	-0.07419	0.02029	-0.16213	35.70193	0.60306
	-3.93893	-3.71047	-0.19103	1.22593	-3.87678	1.56230	-2.69816	3.46431	
	-0.36336	-0.34485	-0.01891	0.12050	-0.35836	0.15287	-0.25811	0.32446	
1960 H	-8.13806	-0.00260	0.03290	0.02898	-0.07145	0.01648	-0.10442	36.83440	0.57899
	-5.23768	-4.30426	0.59833	0.24211	-3.19510	1.18716	-1.29624	4.40581	
	-0.46038	-0.39206	0.05914	0.02397	-0.30163	0.11674	-0.12730	0.39985	
1961 H	-7.44128	-0.00220	-0.02445	0.26000	-0.05860	0.04050	-0.10502	36.09426	0.52093
	-3.39297	-2.66846	-0.43248	2.01559	-2.42090	1.30952	-1.10619	2.86370	
	-0.31846	-0.25545	-0.04278	0.19571	-0.23310	0.12858	-0.10888	0.27279	
1962 H	-8.30935	-0.00046	0.02129	0.37139	-0.06588	0.04752	-0.08924	36.20368	0.45511
	-3.51066	-2.77626	0.34089	2.42047	-2.51657	1.44375	-1.02624	3.12577	
	-0.32834	-0.26506	0.03373	0.23306	-0.24178	0.14151	-0.10109	0.29566	
1963 H	-0.68543	-0.00074	-0.05717	0.13498	-0.05616	0.05070	-0.15260	-0.85236	0.52089
	-0.28807	-4.10707	-1.37450	1.20525	-2.87385	1.56855	-1.71624	-0.08817	
	-0.02851	-0.37670	-0.13485	0.11850	-0.27369	0.15347	-0.16753	-0.00873	
1964 H	-0.79867	-0.00100	-0.08541	0.09822	-0.08998	0.04940	-0.23655	0.09914	0.59187
	-0.60014	-4.60927	-2.12253	0.89933	-5.01308	1.89804	-2.69183	0.01522	
	-0.05932	-0.41519	-0.20567	0.08870	-0.44461	0.18470	-0.25754	0.00151	
SET 4.R	X2	X3 AER	LOG X4	LOG X5	LX6AER	X7 AER	X5 M3	X2SP	
1958 R	-7.43694	-0.00221	0.09652	-0.11059	-0.09293	0.01782	-0.16356	30.13419	0.56976
	-5.42773	-4.17236	1.68025	-1.08535	-4.69865	0.77602	-2.43641	4.82889	
	-0.47339	-0.38182	0.16411	-0.10685	-0.42182	0.07661	-0.23451	0.43136	
1959 R	-8.87027	-0.00202	-0.01931	0.11349	-0.07436	0.01913	-0.17116	41.35880	0.59954
	-4.46326	-3.69462	-0.40596	1.07076	-3.82100	1.69171	-2.82522	4.00255	
	-0.40422	-0.34356	-0.04016	0.10543	-0.35386	0.16520	-0.26940	0.36843	
1960 R	-8.64392	-0.00205	0.04244	0.03732	-0.07179	0.01549	-0.10315	40.34858	0.54675
	-5.37452	-3.27753	0.74849	0.30064	-3.07034	1.26210	-1.23117	4.67775	
	-0.46978	-0.30868	0.07391	0.02975	-0.29086	0.12400	-0.12101	0.42028	
1961 R	-7.55346	-0.00141	-0.01500	0.20498	-0.05677	0.01659	-0.11483	37.03005	0.49714
	-3.36028	-1.86370	-0.26216	1.55338	-2.30227	0.35921	-1.14616	2.86333	
	-0.31570	-0.18147	-0.02595	0.15202	-0.22226	0.03555	-0.11276	0.27276	
1962 R	-7.85739	-0.00076	0.01189	0.38867	-0.05869	0.06454	-0.07739	34.09316	0.45425
	-3.32862	-2.84745	0.19232	2.53343	-2.30292	1.43132	-0.88451	2.95026	
	-0.31302	-0.27136	0.01904	0.24331	-0.22232	0.14032	-0.08725	0.28040	
1963 R	-0.49081	-0.00049	-0.06116	0.19802	-0.05258	0.12967	-0.12935	-0.99354	0.48681
	-0.19920	-2.70098	-1.44655	1.72834	-2.68474	2.58511	-1.41217	-0.09942	
	-0.01972	-0.25836	-0.14178	0.16868	-0.25691	0.24797	-0.13848	-0.00984	
1964 R	-1.21101	-0.00046	-0.09147	0.14400	-0.09136	0.06998	-0.20194	2.87895	0.55811
	-0.86822	-3.39878	-2.18825	1.27789	-4.94705	2.20977	-2.22974	0.42096	
	-0.08565	-0.31895	-0.21176	0.12553	-0.43989	0.21374	-0.21559	0.04165	

MINUTES OF THE ANNUAL MEETING OF THE SOCIAL STATISTICS SECTION

Pittsburgh, Pennsylvania, August 21, 1968

The meeting was open by John D. Durand, Chairman, at 5:30 p.m. The Chairman announced that, as a result of the recent elections, the full listing of officers for 1969 was:

Chairman	Henry S. Shryock
Chairman-Elect	Daniel O. Price
Vice-Chairman	Eleanor B. Sheldon
Vice-Chairman	Eva Mueller
Secretary	Regina Loewenstein
Publications Liaison Officer	Denis F. Johnston
Section Representative on the Board of Directors	Margaret E. Martin
Section Representative on the Council	Leslie Kish
Editor of Proceedings	Edwin D. Goldfield

Members were encouraged to submit papers of substantive interest for publication in the Journal. The Journal embraces statistical applications as well as theory and method, but the Editor receives few papers on applications, according to a report to the meeting by Margaret Martin.

The encouragement of such papers should be a major task of the Publications Liaison Officer. It was suggested that he might scrutinize the Proceedings for potential Journal articles, and raise the question of eliciting papers on applications which might qualify for publication in the Journal with the Section program chairman and the session chairmen when sessions for the annual meetings are being planned. A motion was duly made, seconded, and adopted, making the term of the Publications Liaison Officer two years.

The Chairman invited comment on the social statistics sessions in the 1968 meetings. Greater attention to crime statistics was urged. The presence of one session on this topic in the 1968 program was noted with approval, and the need for more was asserted.

The question was raised, whether the "data bank" issue was a live issue, and if so, whether the Section should propose a watchdog committee to keep posted on it. It was pointed out in response, that any Federal data bank would require legislation. If there should be any draft legislation by the time of the next meeting, a session might be scheduled to discuss it.

It was suggested that in five to ten years the problems of cities would require a great evolution in the number of variety of social statistics, and that there would be a need to develop Social Indicators as an aid to policy formation in this sphere. The Chairman suggested that the program committee might consider the possibility of a session on this subject being held at the next annual meeting.

Proposed legislation limiting the scope and the compulsory character of the Census was noted with concern. Although no action is expected in this session of Congress on the 43 bills in the House and one in the Senate, reintroduction of the bills and more hearings are expected in the next session. Representative to the Board of Directors was asked to express the concern of the Section to the joint meeting of the Board and Council. The American Statistician has carried stories describing this situation. It was reported that the American Public Health Association sent its members a notification of the problem. A suggestion was made that the Association send a letter to the members on this problem. It was also suggested that any statement by the Association, including any letter to the members, should be limited to a description of the situation.

The Chairman noted that the time of this meeting was selected pursuant to the entreaty of the Section at its December 1967 meeting, that the time should be later than breakfast. A motion was duly made, seconded, and adopted, that the next meeting should convene at the same time as this meeting.

Because the Secretary was unable to attend, these minutes were taken by Robert Parke, Jr.

1968 Officers of the Social Statistics Section

Chairman	John D. Durand
Chairman-Elect	Henry S. Shryock
Vice-Chairman	Elijah L. White
Vice-Chairman	Eleanor B. Sheldon
Secretary	Regina Loewenstein
Representative on Board of Directors	Margaret E. Martin
Council Representative	Leslie Kish
Editor of Proceedings	Edwin D. Goldfield