SUBSTITUTION FOR MISSING UNITS1/

David W. Chapman, Bureau of the Census

1. Introduction

In nearly every sample survey, especially those that involve human population, there will be some eligible units selected into the sample for which all or part of the survey data items are not obtained. Eligible sample units for which no usable survey data are collected are referred to as total nonresponse cases (or total questionnaire nonresponse cases or unit nonresponse cases). Sample units for which only a part of the data items are obtained are referred to as partial (or item nonresponse cases.

Whenever there is nonresponse in a survey, there will necessarily be some bias in the survey estimates. There is no method of correcting nonresponse bias since the missing survey characteristics of the nonrespondents are, by definition, not available. In order to hold this bias to a minimum level, the data collection phase of a survey should include extensive follow-up procedures to provide a high survey response rate (e.g. 85 percent or more) and a low item nonresponse rate. In order to minimize the bias associated with the nonresponse that remains after the follow-up procedures of a survey are completed, a number of methods of adjusting or imputing for survey nonresponse have been suggested and used. Discussions of the various methods of imputing for both total and partial nonresponse are given by Chapman (1976) and by Bailar, Bailey and Corby (1978). A discussion of the problem of total nonresponse is presented by Kish (1965, pp.532562) and by Platek, Singh, and Tremblay (1978).

Perhaps the most common method of imputing for total nonresponse in a survey is to adjust (upward) the weights of the respondents in a way that accounts for the nonrespondents. Such adjustments are usually made separately within each one of a set of mutually exclusive and exhaustive sample groups, referred to as nonresponse weight adjustment classes. The weights of the respondent in each class are increased by the factor that will cause the sum of the adjusted weights of the respondents to be equal to the sum of the unadjusted weights of all eligible sample cases in the class. Effectively, this procedure imputs (or substitutes) the average values of the survey items of the respondents in each class for those of the nonrespondents in the class. Consequently an attempt is made to define weight adjustment classes in such a way that the respondents and nonrespondents in a given class will have similar survey characteristics.

Another procedure that is used in many surveys to impute for total nonresponse is substitution (or field substitution). Rather than adjusting the weights of the respondents to impute for nonrespondents, as is done in weight adjustment procedures, population units not originally selected for the sample are used to replace place eligible sample units that do not participate in the survey. In designing a substitution procedure, an attempt is generally made to arrange for substitutes that will have

The remainder of this paper is devoted to a discussion of some substitution procedures including an investigation of the impact of substitution on survey estimates.

2. Methods of Substitution for Nonrespondents

In general, two basic types of substitution procedure are used:

(1) Selection of a random substitute
(2) Selection of a specially designated substitute.

With a random substitution procedure, an additional population unit is selected on a probability basis to replace each nonrespondent. Usually the substitute for a particular nonrespondent is chosen from a restricted population subgroup (e.g. the same block, enumeration district, stratum, or group of strata from which the nonrespondent was selected). In such cases it is hoped that the characteristics of a substitute selected from a population subgroup will be more nearly like those of the nonrespondent than would the characteristics of a substitute selected from the entire population.

For many random substitution procedures, potential substitutes (i.e., alternative or backup units) are selected by a random procedure prior to the data collection phase of the survey. This avoids any delay and trouble that would be involved in selecting a substitute for a nonresponding sample unit after the data collection activities have begun. Also more than one backup unit is usually selected for each sample unit to allow for nonresponding substitutes.

A procedure that uses specially designated substitute units identifies one or more purposively selected backup units to proivde substitutes, if necessary, for each sample unit. The intent, of coerse, is to specify substitute units that have characteristics similar to those of the nonrespondents (e.g., a geographic neighbor of a nonrespondent or a unit that has specified characteristics identical with or similar to those of the sample unit.)

When a substitution procedure is used, there will generally be some survey nonrespondents for which substitutes are not obtained. In such cases, substitution should be used in combination with another imputation procedure to account for all the nonrespondents.

Example 1. An example of a random substitution procedure is one used in the National Longitudinal Study (NLS), sponsored by the National Center for Educational Statistics. This procedure, which is discussed in more detail in Section 4.1, is described in a report by Williams and Folsom (1977). The first stage of the NLS sample design was the selection of a probability sample of four secondary schools from each of 600 strata. Two of the four schools were randomly selected for the primary sample, while the other two were designated as backups (i.e., possible substitutes). If either one or both of the primary schools did not participate in the NLS, one or both of the backup schools were approached, as needed, to be sample substitutes. If attempts to obtain cooperation from the two

backup selections in a given stratum failed, no other substitutes were used. In such cases weight adjustments were made to impute for nonresponding schools.

Example 2. An example of a substitution procedure using specially designated substitutes is one used in the Michigan Survey of Substance Use, described by Sirken (1975). The sample design for this survey, which is discussed in more detail later in Section 4.3 was a multistage household sample of 2,100 households. If the designated interviews were not completed after three calls at a household, the household was dropped from the sample and an attempt was made to obtain a substitute household. In total there were 700 nonrespondent households in the survey. The first household approached to be the substitute was the one directly to the right of the nonresponding household. If this attempt failed, the household to the left of the nonrespondent was approached. If this attempt also failed, other households were approached in some manner (not descirbed by Sirken) until a substitute household was obtained.

Example 3. An example of a substitution scheme involving both of the basic types of substitution procedures is one designed by Westat Research, as a subcontractor to the Educational Testing Service (1972, Chaper 3) for a study carried out for the National Center for Educational Statistics. The design for this study involved the selection of a systematic sample of two schools, with probability proportional to a measure of size, in each one of about 500 strata. At the same time the two schools were selecded, a systematic sample of eight additional schools was selected per stratum to be used as backup schools. The first four selected were designated as backups for one of the primary selections and the other four were designated as backups for the other primary selection. If a primary school declined to participate after numerous requests, a substitute was approached. (The school nonresponse rate was about about 10%). The first priority substitute was that school, if any, that (1) was located in the same school district, (2) was placed in the same stratum (3) had the same grade structure, and (4) had the closest enrollment size to that of the nonrespondent school. If this attempt failed, the four systematically selected backup schools were approached in a pre-designated sequence. With this procedure the top priority backup school is a designated substitute while the other backup schools are randomly selected schools. Using this substitution procedure, which was implemented by the Educational Testing Service's Berkeley, California office, a substitute school was obtained for nearly every nonrespondent school.

3. Advantages and Disadvantages of Substitution Procedures.

A criticism of substitution procedures that is often made--see, for example, Kish (1965, p 558)--is that substitution is of no help in reducing nonresponse bias since the nonresponses are replaced by responses that are presumably like the responses already in the sample. This is an unfair criticism if directed solely at the use of substitution procedures since all the methods used for nonresponse imputation, including weight adjustment procedures, suffer from that same basic weakness: data for nonrespondents have to be supplied (imputed) from data provided by respondents. The key question regarding the worth of substitution procedures is whether or not the use of substitutes provides better proxy values for nonrespondents than those provided by alternative imputation procedures.

A specially designated substitution procedure often uses a substitute that is (1) a neighbor of the nonrespondent, or (2) a listing that is adjacent to the nonrespondent on the sampling frame. Because of this proximity, this type of substitution procedure might provide better proxy values than would be obtained by using an imputation procedure that involves some type of "averaging" of the characteristics of a subclass of respondents. However, in cases in which some information is available about the nonrespondents, better proxy values might be available from sample respondents, than from substitutes, since much more information is available about respondents than about potential substitutes.

For example, if the race of a nonrespondent household is obtained, this information could easily be used for an imputation procedure that is based on the characteristics of respondents. However, race information cannot generally be used in obtaining substitutes in a household survey.

Perhaps the type of survey for which the use of substitution is most appropriate is one that involves a deeply stratified, relatively small sample of population units.2/ Surveys of institutions (e.g., schools or hospitals) are often of this type since a substantial amount of stratification information may be available for a population of institutions, and the cost per sample unit might be high enough to force the sample size to a minimum level. In such cases, a substitute for a nonrespondent might be available that is very similar to the nonrespondent with respect to a number of characteristics. However, for the type of weight adjustment procedure often used, the weight adjustment classes would generally have to be large enough (to avoid producing a nontrivial increase in survey variances) to include institutions with substantial variation in some characteristics. Consequently, in surveys of this type, the use of substitute institutions would tend to provide better imputations than would the type of weight adjustment procedure that is often used.3/

For surveys with large sample sizes and relatively little information about nonresponding units (e.g., a large household survey), the use of substitutes would probably not provide any improvement in terms of bias reduction over the use of weight adjustment procedures. However, some reduction in survey variances would result due to an increase in sample size. In such cases the small reduction in survey variances may not be worthwhile if one considers the increase in survey costs and the potential problems that could result (see the disadvantages of substitution procedures described below).

In addition to the bias reduction that may or may not result from the use of substitution in certain situations, as discussed above, there are

two advantages and three disadvantages that generally apply to substitution procedures. The first advantage is that the sample will be balanced with respect to sample size per substitution class. This balance has certain practical advantages. For example, if a self-weighting sample is selected, the final sample will still be essentially self-weighting, since the most appropriate selection probability to use for a substitute unit is usually the probability that the unit had of being selected into the initial sample. Furthermore, if a two-perstratum design were used in order to allow for variance estimation, there would still be two units per stratum if substitutes were selected from the same stratum.

The other advantage to the use of substitution, which has already been mentioned, is that, for a fixed initial sample size, it increases the survey sample size and therefore reduces the variances of survey estimates. Of course, this increase in the sample size generally, involves an increase in survey costs, which is the most obvious disadvantage of using substitution.4/

Perhaps the major disadvantage of the use of substitution is that the effort extended to obtain participation from originally selected units may not be as intense as it would if no substitutes were available. That is, an interviewer, and perhaps a research analyst, may view a backup unit as one that is just as good (or nearly as good) as the unit initially selected. This could lead to a higher level of nonresponse which could produce greater biases in the survey results. If a substitution procedure is used, steps should be taken in the planning of the data collection procedures to ensure that the maximum effort is made to obtain responses from the original sample units. This should include validating a substantial proportion of the substitute units to verify that substitutes were needed.

The third disadvantage of the use of substitutes is that there is a tendency to ignore the level of substitution used when the survey response rate is reported. That is, since survey practitioners process substitute responses when survey estimates are computed as if the substitutes were selected in the original sample, there there is a tendency to view the substitutes as originally selected units when response rates are calculated. In fact, in some surveys, a record of whether or not a respondent is a substitute is not easy to find. If the substitutes are treated as original sample selections for the calculation of the response rate, the survey response rate will, of course, be overestimated, and the potential for nonresponse bias will be underestimated. Whenever substitutes are used in a survey, care should be taken to (1) keep accurate records of which units are substitutes, (2) identify which data records are obtained from substitute units, (3) report the level of substitution, and (4) treat the substitutes as nonresponse cases when calculating the survey response rate.

4. Summaries of Some Research Investigations of the Impact of Substitution on Survey Estimates

An extensive search of the survey research literature, both published and unpublished, has not uncovered any theoretical work relating to substitution. Perhaps this is because any model that is used to represent a substitution procedure would either be too complex to formulate or too simple to provide useful results. Specifically, a model used to represent a specially designated substitution procedure would have to be complex to appropriately reflect that type of substitution process. It would be very difficult to adequately reflect the relationship between the characteristics of the substituttes and those those of the nonrespondents or between those of the substitutes and those of the respondents.

On the other hand, a model representing a random substitution procedure within post strata would be simple since the random substitutes would have the same expected characteristics as the respondents in the post stratum from which substitutes are selected, although not necessarily the expected characteristics of nonresponding units. For example, if simple random sampling is used to select the initial sample and the substitutes from a stratum, the expected value of the resulting stratum sample mean would be the stratum population mean for respondents. The bias of this estimate is the difference between the stratum population mean for respondents and the overall stratum population mean. The bias of the stratum sample mean would be the same if a weight adjustment procedure were used, assuming that the stratum would be used as a weight adjustment class. (However, the survey variances associated with the weight adjustment procedure would be somewhat larger.)

For a random substitution procedure to provide estimates with lower nonresponse bias than those that would be provided by a weight adjustment procedure, the substitution classes would have to be defined differently--presumably narrower--than would the weight adjustment classes. These classes would have to be defined in such a way that the differences between the population means for respondents and nonrespondents within a class would be smaller on the average than the corresponding differences within a weight adjustment class.

It was also difficult to find research involving empirical studies of the impact on survey estimates of substitution procedures. In particular, only four studies were discovered that provided estimates of the bias associated with one or more substitution procedures. Furthermore, only one of these studies included a comparison of the estimated effects (or biases) for a substitution procedure and the corresponding estimated effects (or biases) for one or more alternate imputation procedures. Unfortunately, the type of substitution procedure used in the comparative analysis was atypical and the scope of comparisons was rather limited.

The four studies mentioned above are summarized in the remainder of this section, with the comparative study being discussed last.

4.1 Study by Reuben Cohen (1955)

The primary objective of Cohen's study was to estimate the bias associated with a substitution procedure that he refers to as a "modified probability sample." Although his procedure could be used for other types of surveys, it is investigated by Cohen in the context of a house-

hold survey. To use Cohen's substitution procedure, the basic household sample has to be selected in pairs of housing units--presumably pairs of neighbors. The number of pairs selected is equal to the target sample size. One unit of each pair of units is selected at random to be the primary selection while the other unit is designated as the alternate or backup selection. A call is made first at the primary unit in an attempt to obtain an interview. If an interview is obtained, the interviewer proceeds to the next pair. If an interview is not obtained on the first call, (s)he attempts to obtain an interview with the alternate unit. If an interview is completed with the alternate unit, the interviewer continues to the next pair. If an interview is not obtained, (s)he returns to the primary unit, then back to the alternate, continuing in this manner until an interview is obtained.

Cohen wanted to find out whether this modified probability sample, which would reduce field costs, would introduce a nontrivial amount of bias in the survey estimates.

Cohen based his investigation on a subset of data collected in a household survey conducted by the Bureau of Social Science Research (BSSR). A brief description of the BSSR survey will be given, followed by a description of how Cohen carried out his study using a subset of the BSSR sample.

The BSSR study was a household survey carried out in the Washington, D.C., metropolitan area. The purpose of the BSSR survey was to study the effects on the community of the purchase of the Times-Herald by the Washington Post. The BSSR selected an equally weighted, multi-stage sample of about 550 pairs of adjacent households.5/ From each sample household an attempt was made to interview one person aged 18 or more, selected randomly from the eligible household members.

To investigate the modified probability sample substitution procedure, Cohen could only use those pairs of households in the BSSR survey for which completed interviews were obtained from members of both households. For this restricted subset of the sample, which included 326 of the original 550 pairs, Cohen was able to simulate how the sample would be selected using the modified probability sample. This was possible since a record of the number of calls required to obtain an interview was recorded for the BSSR survey.

Cohen randomly designated one household in each of the 326 pairs to be the primary unit. Based on the number of calls required to obtain an interview for each of the two households in a pair, Cohen was able to determine which of the two households would be included in the "modified probability sample." Comparisons were then made between the primary probability sample of 326 households and the modified probability sample of 326 households. These comparisons were done in the second part of his analysis.

In the first part of his analysis he did paired comparisons tests to compare the early and late respondents with respect to 12 survey items (all 0-1 items). These tests were carried out only for the 209 pairs (of the 326 pairs) for which one of the units responded on an earlier call than the other. For the 12 items included Cohen found significant differences in the sample proportions for only two items:

(1) Proportion not in the labor force (higher for early responders)
(2) Proportion who reported reading the Washington Daily News (higher for late responders)

There were two other items for which the difference in sample proportions was almost significant at the 10% level:

(1) Proportion that are heads of household (higher for late responders)
(2) Proportion residing in households with heads who are professionally or technically employed (higher for early responders)

For these two paired samples none of the other eight sample proportions were significantly different at the 20% level. These proportions included for example, the proportion with one or more children in a household, the proportion male, the proportion with family income over $5,000, and the proportion under 40 years of age.

The second part of his analysis involved calculating an estimate of bias attributable to use of his substitution procedure. For the set of 326 interview pairs he calculated the unweighted sample proportions for twelve survey items using a "straight" probability sample (i.e., the 326 primary units) and also using the modified probability sample of 326 primary and backup units. The estimated proportion reading the Washington Daily News was about 0.03 lower with the substitution procedure. For the other ten survey items, the estimated proportions using the two procedures were within 0.02 of each other.

It should be noted that Cohen was only able to compare primary and substitute units for those cases in which both members of a selected pair of households supplied completed interviews. This restricted population may provide different comparisons between primary and substitute units than would be provided by the entire target population. In particular, no comparisons can be made in his study of the characteristics of the substitutes and the characteristics of the nonresponding primary units being replaced.

Furthermore, the substitution procedure investigated by Cohen differs from those typically used. With most procedures, in order to minimize the nonresponse bias, more than one call is generally made at the primary unit before a backup unit is approached. In fact, as was indicated earlier in Section 3, a strong effort to obtain an interview from the primary unit before approaching the substitute unit is highly recommended. The savings on field costs that would be obtained by the use of Cohen's modified probability sample may not be worth the nonresponse bias associated with the additional substitutes used.

4.2 Study by Stephen Williams and Ralph Folsom (1977)

The study by Williams and Folsom of the Research Triangle Institute was done for the National Center for Educational Statistics (NCES). This study was based on data collected in the NCES sponsored National Longitudinal Study of the High School Class of 1972 (NLS). They attempted to measure the portion of bias in the first year estimates that is attributable to school nonresponse.

The basic objective of the NLS was to deter-

mine what happens to students after they graduate from high school. The variables of interest included type and amount of additional education, vocational experience, plans, and attitudes. The NCES wanted to relate this information to the students' personal and educational background. The full-scale NLS was begun in January 1972; and the first phase of the NLS, referred to as the basic year (BY) survey, ended in June 1973. The primary survey instruments for the BY survey were student, counselor, and school questionnaires, student test booklets, and students' school record information forms.

The sample design for the BY survey was a stratified two-stage probability sample with schools as the first stage and students as the second stage. The first stage primary sample consisted of the selection of 1,200 schools--two from each of 600 strata. Two additional schools were selected per stratum on a probability basis to serve as backups (potential substitutes) for the two primary selections. At the second stage, a sample of 18 students was selected from each cooperating school.

If either one or both of the primary schools selected from a stratum declined to participate, one or both of the backup schools were approached. In strata for which the two backup schools selected did not provide the substitutes needed, no additional units were approached. In total, 974 schools participated in the BY survey: 921 primary schools plus 53 substitutes. The student questionnaire was completed by 16,409 students. The nonparticipating primary schools for which substitutes were not obtained were accounted for in the estimation procedures by making adjustments in the weights of the cooperating schools.

From October 1973 through April 1974 the first data collection follow-up procedure (FFU) of the NLS was conducted. In addition to administering a questionnaire to students who participated in the BY survey, there was a complete follow-up of BY nonrespondent schools. A sample of persons who were seniors in 1972 was selected from each of these BY nonrespondent schools. The questionnaire administered to this sample of 1972 seniors included 14 questions about information initially requested in the BY survey. The FFU data gathered retrospectively by BY nonresponding schools allowed Williams and Folsom to make some estimates of bias in the BY survey estimates due to nonresponding schools.

In the FFU survey, 1,300 sample schools completed FFU questionnaires: 1,153 primary schools plus 131 back-up schools plus 16 additional schools selected to account for an increase in the frame.6/ The FFU survey contained completed questionnaires from 21,350 students.

The main purpose of the study by Williams and Folsom was to estimate and analyze the part of the bias in the BY estimates that is attributable to school nonresponse. This included the effect on bias of the use of substitutes and of weight adjustments to account for nonrespondent schools. Other sources of bias in the BY estimates (e.g., student nonresponse in participating schools) were not investigated in their study. Estimates of the biases due to nonresponding schools were made for estimated totals and proportions for 35 student items. These estimates were calculated by comparing the BY estimates obtained from the BY survey data with the "best estimates" computed from the BY data and from the retrospective data obtained in the FFU survey from the sample of 1972 seniors in BY nonrespondent schools. For the BY survey about 20% of the primary schools did not participate. However, after the FFU survey only 2% of the BY primary schools were still nonrespondents. Consequently, differences in the "best estimates" and BY survey estimates should reflect the impact on BY survey estimates of the use of substitutes and weight adjustments to adjust for the 20% of the primary schools that did not participate in the BY survey. Unfortunately, they were not able to separate the impact of the use of substitutes from the impact of the school nonresponse weight adjustments.

The 35 questionnaire items included in the analysis contained a total of 155 response categories (i.e., multiple choice type responses). Estimates of the bias in BY estimates of student totals were calculated for all 155 categories.7/ For each category the null hypothesis that there is no bias in BY estimates was tested against a two-sided alternative. For the 5% level of significance the null hypothesis of no bias was rejected 91 times out of 155. For these estimates the average estimate of school nonresponse bias in BY estimates is approximately a negative five percent.

These estimates of school nonresponse bias (i.e., the differences between the "best estimates" and BY survey estimates) might have been significantly affected by the response and nonresponse biases of the students selected from the schools that participated in the FFU but did not participate in the original BY survey. This seems quite possible because of the time lag of 1 1/2 years or more that occurred between the BY survey and the FFU survey, and the problem with participation and recall that would probably occur after such a time lag.

Although there were some simplifying assumptions and approximations made in estimating the nonresponse biases, Williams and Folsom state on page 8 of their report that "the results are so consistent and dramatic it is doubtful that a more refined analysis would alter the conclusions."

4.3 Study by Monroe Sirken (1975)

Sirken's investigation was based on data collected in the Michigan Survey of Substance Use, August, 1974, sponsored by the Michigan Department of Health.

The Michigan Survey was a statewide, multistage, self-weighting probability sample of the noninstitutionalized population 13 years of age and older, residing in Michigan during August 1974. The main objective of the survey was to estimate the number and proportion of persons in the population and in various subpopulations that used 16 types of substances (drugs) during the previous year.

There were 2,100 households selected for the Michigan Survey. Half of the eligible persons in participating households were interviewed for the survey. If all the designated interviews for a sample household were not obtained after three visits, that sample household was discarded and a substitute household was approached. The prescribed pattern of substitution was to take the household to the right of the nonre-

spondent; then, if necessary, approach the household to the left. Only one call was made to a substitute household. In spite of this, a substitute household was obtained for each nonrespondent household. (Sirken's report did not indicate whether or not it was necessary in some cases to contact more than the two adjacent neighbors to obtain a substitute, and if so, how this was done.)

Of the 2,100 households selected, interviews were completed in 1,400 households. Substitute households were obtained for the 700 nonparticipating households. In total, 2,539 interviews were obtained in the Michigan Survey: 1,552 from initially selected households, 839 from substitute households, and 148 that have unknown status. That is, about one-third of the survey responses came from substitute households.

Sirken compared some survey characteristics by age categories for the following three groups of respondents:

(1) Initially designated individuals that were interviewed on the first call
(2) Initially designated individuals that were interviewed on the second or third call
(3) Substitute individuals

The survey items used for the comparison of the three groups were the percents of persons using (1) alcohol, (2) prescribed drugs, (3) nonprescribed drugs, and (4) illicit drugs in the past year. These comparisons indicated that the characteristics of the substitute persons were about the same as those initially designated persons interviewed on the first call. (This is not surprising since only one call was made at each substitute household.) For the oldest age group included in the analysis, 35 years and older, the characteristics of those persons interviewed on the second or third call were also about the same as for those interviewed on the first call and for those of the substitutes. However, for the youngest age group, 13-17 years of age, and for all four survey items, the substance use rate was higher for the persons interviewed on the second or third call than it was for substitutes and for persons interviewed on the first call. The same relationship held for the middle age group, 18-34 years of age, for two of the four survey items used in the comparison: percent usage rates for alcohol and for illicit drugs.

A similar relationship was observed between early and late respondents for a survey of Vietnam veterams (Wish, et al. (1978)). They found a statistically significant difference between early and late respondents for the proportion of veterans experiencing various types of behavioral problems (e.g., alcohol abuse, arrests, unemployment, and divorce). In all cases the proportion experiencing the problem was higher for the late responders.

Based on the evidence in the Michigan Survey that the substance use rate increases as the number of required calls increases, Sirken suggests that the characteristics of the persons in nonresponding households might be closer to those of persons interviewed on the second or third call than to those of the substitute persons. Sirken shows that if this is the case in the Michigan Survey, there would be a negative bias of about three percentage points in the estimates of the substance use rates for alcohol and illicit drugs for both the 13-17 and 18-34 age groups.

Since the characteristics of the Michigan Survey nonrespondents were not known, there is no basis for evaluating the accuracy of Sirken's speculative estimates of bias. If survey data for at least some of the nonrespondents had been obtained, more objective estimates of bias associated with the substitution procedure could have been made.

4.4 Study by J. Durbin and A. Stuart (1954)

The investigation by Durbin and Stuart was a comparative study of alternative methods of coping with total questionnaire nonresponse. This study, which was carried out in six urban areas in Great Britain, was part of an experimental study planned and directed by a research group called the Survey Research Committee.

The frame (and presumably the population) from which the sample was selected for this study was the Electoral Register. A probability sample of 1,260 names was selected for the sample, 360 from one of the six survey areas and 180 from each of the other five areas. Within each of these six areas, uniform selection probabilities were used. However, the uniform selection probability used to select registrants in one area was, in general, different from that used to select registrants in the other areas. These differential selection probabilities were not taken into account in the comparative analysis. That is, for calculating estimates the entire sample was treated as self-weighting.

The survey data were collected by personal interview. The questionnaire contained items which covered a variety of topics, including age, sex, marital status, employment characteristics, leisure activities, and smoking and drinking habits. The comparative analysis was based on estimated proportions for 32 of the questionnaire items.

There were seven methods of coping with nonresponse that were compared:

(a) and (b)
Make a maximum of (a) one call or (b) three calls and assume that the respondents constitute a probability sample from the total population.

(c) Make an unlimited number of calls and assume that the respondents constitute a probability sample from the total population.

(d) Make only one call and adjust the selection probability of each respondent to take into account the number of days in the previous six days that the respondent was home (i.e., the Politz-Simmons method, described by Politz and Simmons (1949)).

(e), (f), and (g)
Make a maximum of (e) one call, (f) two calls or (g) three calls and obtain substitutes for the nonrespondents using a quota sampling procedure.

The substitution procedure described above in (e)--with variations (f) and (g)--is rather unusual. Instead of selecting substitute individuals from the Electoral Register to replace nonrespondents, the nonrespondents are replaced by a quota sample. Although Durbin and Stuart do not give the details of how the quota sample was sel-

ected, they indicate that the interviews were conducted mostly outside the home, and that the control factors for the quota sample were sex, age, and income class. An attempt was made to obtain the control factors for each sample person not interviewed on the initial call by requesting this information from other household members or, if necessary, from neighbors. The cell counts of the first-call nonrespondents for the three-way cross-classification of the control factors provided the quotas for the quota sampling for procedure (e). Quotas to be used for procedures (f) and (g) were obtained in an analogous way from cell counts made after two and three calls. Using this type of substitution procedure, many of the substitutes used may not even be listed on the sample frame (i.e., the Electoral Register), depending on exactly how the quota sample of replacements was selected.

Of the original sample of 1260 names, 80 were eventually classified as ineligible, leaving an eligible sample size of 1180 persons. The response rates after various numbers of calls are the following:

Number of Calls	Number of Responses	Response Rate
1	374	31.7%
3	823	69.7%
Unlimited	938	79.5&

The above table indicates that even with unlimited callbacks the survey response rate was only 79.5%. Substitutes from the quota sample were obtained for most, but not all, of the nonrespondents. (Durbin and Stuart do not give reasons for the apparent shortage of substitutes.) The number of substitutes obtained for the nonrespondents are given in Table 1 for procedure (e)--substitution after one call--and for procedure (g)--substitution after three calls. The 259 quota substitutes used in procedure (g) consisted of those among the 701 substitutes used in procedure (e) that replaced first-call nonrespondents who remained nonrespondents after three calls. It seems that the total number of interviews for procedures (e) and (g) should have been the same. The discrepancy of seven interviews shown in the above table was not discussed by Durbin and Stuart.

The method of analysis used by Durbin and Stuart was to compare the estimates based on all the procedures to those obtained for the unlimited callback (U.C.) procedure. That is, the estimates based on the U.C. procedure were considered to be unbiased for this experiment, or at least to be the best estimates to use as a standard for comparison. This would have been a sound approach to making the comparisons if the response rate for the U.C. procedure had been closer to 100%. However, since the response rate for the U.C. procedure was only 79.5% the method of comparison is not sound since the estimates based on the U.C. procedure might contain a substantial amount of nonresponse bias.

In particular, the method of comparison to the U.C. estimates is especially unfavorable towards the substitution procedures and the Politz-Simmons procedure. The substitution and Politz-Simmons procedures attempt to impute for all survey nonrespondents, not just for those who eventually become respondents using unlimited callbacks. Consequently, for the 32 questionnaire items included in the comparison analysis, all three substitution procedures and the Politz-Simmons procedure provide estimates that often vary considerably from those of the U.C. procedure. Of these four alternatives, procedure (g)--substitution after three calls--compares most favorably with the U.C. procedure.

There was not much of a pattern observed in the differences between the estimates based on the U.C. procedure and those based on the best substitution procedure (i.e., the procedure for which substitutions were used after three calls). Sometimes the estimated proportions were higher using the U.C. procedure, but other times they were higher for the three-call substitution procedure. It did appear, however, that the quota sample substitutes used after three calls constituted a somewhat younger, less affluent group than did the respondents
who cooperated after three calls.

Of all six procedures being compared to the U.C. procedure in the comparative analysis (i.e., (a), (b), (d), (e), (f), and (g)) the one that provided estimates closest to those of the U.C. procedure was procedure (b) making a maximum of three calls and assuming the respondents constitute a probabililty sample from the total population. The favorable appearance of the three-call procedure is not surprising since the three-call sample makes up 80% of the U.C. sample (i.e., 823 of the 938 U.C. respondents).

If the U.C. procedure has to be taken as a standard of comparison, then all the procedures compared should be geared to that level of response. That is, all the procedures should attempt to impute to the sample of 938 U.C. respondents and not to the total eligible sample of 1180 persons. If this had been done, the substitution procedures might have compared more favorably with the other procedures than they did.

Even though the substitution procedures included in the study may be somewhat better than they appeared in this experiment, it does not seem that the generation of substitutes from a quota sample is a very promising imputation method. This type of procedure could be difficult to apply to surveys of institutions since it might be hard to obtain control factor informationn for quota substitutes. Also, it would be difficult to use in household surveys since the quota control factors for nonrespondents might be difficult to obtain in some surveys. Even if control factors could be obtained for household nonrespondents, it is questionable whether, in general, a replacement obtained from a quota sample would provide a better substitute than would a neighbor.

5. Conclusion

As was pointed out in the beginning of Section 4, there does not appear, at this time, to be any theoretical results available that would be helpful in evaluating the general usefulness of substitution procedures as a method for imputing for total questionnaire nonresponse. There appear to be some situations for which the use of a substitution procedure would be appropriate, and other situations for which such a procedure would not be wise.

Perhaps the only way that substitution procedures can be evaluated is through empirical investigations. Only four studies that investigated the impact of substitution on survey estimates have been uncovered. Although none of these studies was carried out under ideal conditions, they all seemed to indicate that substitution procedures do not eliminate the effects of nonresponse bias. This was perhaps best demonstrated in the Research Triangle Study for the National Center for Educational Statistics, described by Williams and Folsom (1977). This study indicated that there was consistently a negative 5% bias in a number of estimated totals that was attributable, to a large extent, to the use of substitute schools and to school nonresponse weight adjustments. The fact that substitution procedures do not seem to eliminate nonresponse bias does not imply that substitution procedures are inappropriate. It is probably true that there is no procedure available that can adequately correct nonresponse bias. In a study carried out by Westat Research for the National Center for Health Statistics, described by Chapman (1974), it was discovered that none of the six alternative weight adjustment procedures that were compared could correct the nonresponse bias associated with a number of items in a medical history questionnaire.

To evaluate the usefulness of substitution procedures, empirical studies should be carried out that compare imputations based on substitutes with those based on alternative nonresponse procedures. In an ideal study, sample units would be declared nonrespondents after a specific number of calls (e.g., three). The substitution procedure being evaluated would then be applied to obtain substitutes for the nonrespondents. However, extensive efforts would continue to obtain cooperation from the nonrespondents. The data collection activities would proceed until all of the originally designated sample persons or institutions, except the hard-core refusals, were persuaded to participate in the survey. Estimates based on the substitutes could then be compared to estimates based on alternative weight adjustment procedures with respect to how close the estimates are to those that are nearly free of nonresponse bias (i.e., those based on almost all of the sample).

A study of this type is being proposed for FY 1983 as part of the research being carried out by the Bureau of the Census to investigate the potential use of random digit dialing (RDD) for demographic surveys. For an RDD survey, or for any telephone survey, it would generally be easier to devise, execute, and control a substitution procedure than it is for an in-person survey. Consequently, there is interest in developing and evaluating a substitution procedure as part of the RDD research at the Census Bureau. Basically, the proposed research involves the application of a procedure to generate substitute telephone residences for telephone residences in the sample that do not respond after various numbers of attempts. The experimental survey would also include an extended follow-up effort on the nonresponse cases, permitting some comparisons of the characteristics of nonrespondents with those of their substitutes. It would also be possible to derive and compare estimates that use substitutes to account for nonresponse to those that use nonresponse weight adjustments.

Table 1

Substitution Procedure	Number of Respondents	Number of Nonrespondents	Number of Substitutes	Total Number of Interviews
(e) Substitution after one call	374	806	701	1,075
(g) Substitution after three calls	823	357	259	1,082

Footnotes

1/ Except for some minor editing and updating, this is a paper that was prepared for the Panel on Incomplete Data, Committee on National Statistics, National Academy of Sciences. The Panel report volumes will be published by Academic Press, Inc. (New York, NY).

2/The potential advantage of the use of substitution in this situation was described to David W. Chapman by Joseph Waksberg of Westat Research during a telephone conversation in January 1979.

3/If a raking ratio adjustment procedure were used, the problem of requiring large weight adjustment classes would be avoided. For a discussion of raking ratio estimation, including an extensive reference list, see Oh, H. Lock and Fritz Scheuren (1978).

4/ Since a cost-variance trade-off is difficult to assess, perhaps the best approach for comparing a substitution procedure with a weight adjustment procedure is to set equal the expected number of completed interviews for the two procedures, allowing for a reasonable level of nonresponse. If this were done, the initial sample size associated with the weight adjustment procedure would be larger than that associated with the substitution procedure. However, assuming the anticipated response rate was reasonably accurate, the two nonresponse procedures would involve roughly the same survey costs and variances.

5/ The sample was selected in pairs as a means of providing a mechanism to double the sampling rate for a special segment of the population for certain questionnaire items. This did not affect Cohen's study, however, since his analysis was based only on the questionnaire items asked of all persons in the survey.

6/ The 1,153 primary schools included those that participated in both the BY and FFU surveys and those that participated only in the FFU survey. The 131 backup schools included backup schools used as substitutes for both the BY and FFU surveys.

7/ An estimated total refers to the estimated number of students in the population who would respond in that category.

References

1. Bailar, Barbara A., Bailey, Leroy, and Corby, Carol (1978), "A Comparison of Some Adjustment and Weighting Procedures for Survey Data," Survey Sampling and Measurement, Edited by N. Krishnan Namboodiri, Academic Press, New York, San Francisco, London, 1978.

2. Chapman, David W. (1974), "An Investigation of Nonresponse Imputation Procedures for the Health and Nutrition Examination Survey," prepared for the Division of Health Examination Statistics, National Center for Health Statistics; HEW, by Westat Research.

3. Chapman, David W. (1976), "A Survey of Nonresponse Imputation Procedures," 1976 Proceedings of the Social Statistics Section, American Statistical Association, Part I, pp. 245-251.

4. Cohen, Reuben (1955), "An Investigation of Modified Probability Sampling Procedures in Interview Surveys." A Master of Arts thesis submitted to the graduate faculty of The American University, May 26, 1955.

5. Durbin, J. and Stuart, A. (1974), "Callbacks and Clustering in Sample Surveys: An Experimental Study," Journal of the Royal Statistical Society, Series A (General), Part IV, 1974, pp. 387-410.

6. Educational Testing Service (1972), "Anchor Test Study," Final Report. Prepared for the U.S. Office of Education, National Center for Educational Statistics. Chapter 3 (Sample Design) was prepared by Westat Research.

7. Kish, Leslie (1965), Survey Sampling, New York: John Wiley and Sons., Inc.

8. Oh, H. Lock, Scheuren, Fritz (1978), "Some Unresolved Application Issues in Raking Ratio Estimation," 1978 Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 723-728.

9. Platek, R., Singh, M.P., and Tremblay, V. (1978), "Adjustment for Nonresponse in Surveys," Survey Sampling and Measurement, Edited by N. Krishnan Namboodiri, Academic Press, New York, San Francisco, London, 1978.

10. Politz, A., and Simmons, W.R. (1949), "An Attempt to Get the 'Not-atHomes' into the Sample Without Callbacks," Journal of the American Statistical Association, 44, pp. 9-31.

11. Sirken, Monroe (1975), "Evaluation and Critique of Household Sample Surveys of Substance Use," Chapter IV of "Alcohol and Other Drug Use and Abuse in the State of Michigan," Final Report. Prepared by the Office of Substance Abuse Services, Michigan Department of Public Health.

12. Williams, Stephen R. and Folsom, Ralph E., Jr. (1977), "Bias Resulting From School Nonresponse: Methodology and Findings." Prepared by the Research Triangle Institute for the National Center for Educational Statistics.

13. Wish, E.D., Robins, L.N., Helzer, J.E., Hesselbrock, M., and Davis, D.H. (1978), "Monday Morning Quarterbacking on Limiting Call-Backs: Evidence from a Panel Study of Veterans." Presented at the Annual Meeting of the American Association of Public Opinion Research, Roanoke, Virginia, June 1978.