AN INVESTIGATION OF SUBSTITUTION FOR AN RDD SURVEY

David W. Chapman and Anthony M. Roman, U.S. Bureau of the Census

1. Introduction

In late 1982, the Bureau of the Census and the National Center for Health Statistics formed the Joint Agency Telephone Survey Task Force to plan a three-year program of research and development leading to the implementation of random-digit-dialing (RDD) sampling techniques (via a dual frame design) in the National Health Interview Survey (NHIS). In their final report and three year plan, the Task Force recommended that a feasibility study be conducted early in 1984 to investigate a number of major issues involving the use of RDD in the NHIS. Subsequently, the 1984 NHIS/RDD Feasibility Study was conducted from late January to May. The sample for the study consisted of about 1500 telephone households for each of two questionnaire versions.

One of the objectives of the Feasibility Study was to develop and test nonresponse adjustment procedures. The method that is probably used most often to impute for unit nonresponse in surveys is to adjust (upward) the weights of the respondents to account for the nonrespondents. These adjustments are usually made separately within nonresponse weight adjustment cells. Effectively, this procedure imputes for the survey items of the nonrespondents in each cell the average values of the survey items of the respondents in the cell. An attempt is made to define weight adjustment cells in such a way that the respondents and non-respondents in a cell have similar survey characteristics. To the extent that this goal is accomplished, nonresponse bias will be reduced.

Another method of accounting for unit nonresponse is substitution: replacing a nonrespondent with a unit not originally selected for the sample. The goal in using substitutes is to generate them in such a way that they have characteristics similar to those of the nonrespondents they represent. With respect to calling and interviewing, a substitute is treated the same as an original selection. It is important to identify all substitute cases in the respondent file so that the response rate, based on the original selections, can be calculated.

A major criticism of using substitution is that a substitute might be viewed as being as good, or nearly as good, as the originally selected unit. If so, a reduced effort might be extended to obtain a response from the original unit and substitutes might not be carefully identified in the respondent file. However, with the control over the sampling operation that exists with a centralized RDD-CATI system, these potential problems can be eliminated.[1] Since the interview procedures for substitutes are the same as those for the original sample cases, interviewers would not know whether they were dealing with an original case or a substitute.

Because of the control associated with RDD/CATI interviewing, it was decided to develop and test a substitution procedure for the Feasibility Study. The procedure was evaluated and compared to a weight adjustment procedure.

2. Sample Design for the Feasibility Study.

The sample for the Feasibility Study was selected using the RDD method described by Waksberg (1978). A brief description of how this method was used in this study follows.

Using the telephone exchange file from AT&T, a list of telephone area codes and working three-digit prefixes was created. To these six-digit combinations, all choices of the next two digits were added, forming a frame of the first eight digits in telephone numbers. The eight-digit numbers were the primary sampling units (PSUs). Each PSU contains 100 ten-digit numbers, identified by varying the last two digits. A random selection was made of an eight-digit number (a PSU) and of the last two digits. The number selected was dialed. If the number served a residence, the PSU was labeled "residential" and was retained for the sample. Otherwise, the PSU was labeled "nonresidential" and was excluded. This procedure, referred to as primary screening, was repeated until a specific number, m, of residential PSUs was selected. For each PSU chosen for the sample, additional last two digits were randomly selected and dialed until a specified number, k, of residential telephones was identified for the sample. The process of selecting and attempting to interview k residences in each PSU is referred to as secondary screening. The total sample size for this design is mk.

The Feasibility Study sample was selected in 12 independent replicates.[2] One replicate was introduced each week for 12 consecutive weeks. Each replicate was interviewed for three weeks. The total sample size for the study was about 3,000 telephone residences with a sample size per replicate of about 250. Based on the optimum cluster size formula given by Waksberg (1978), the optimum cluster size for NHIS was estimated to be 6. Also, it was decided to use the same PSUs for the half of the sample assigned to one questionnaire version as for the half assigned to the other version. Therefore, the total cluster size for each PSU was k=12 (six for each questionnaire version). This dictated that m=21 PSUs be selected per replicate to provide about 250 telephone residences. Additional details of the sampling procedures are provided by Tegels and Chapman (1984).

3. Description of the Substitution Procedure

For each residential number selected during secondary screening, an attempt was made to obtain an interview. For those cases that were refusals, other noninterviews, or were numbers which could not be contacted but were identified by a telephone business office as working, substitutes were selected randomly from the same PSU. For a case selected during the initial interview week of the three-week collection period for a replicate, a substitute was selected after the second refusal or after 10 attempted calls to a working number with no contact. For a case selected during either the second or third interview week, a substitute was selected after the first refusal or after 7 attempted

calls to a working number with no contact.

After a substitute was selected, calls were still made to the original sample unit as part of a followup procedure. For refusals, one or two additional calls were usually made. For hard to reach cases, up to 20 calls were made before the case was classified as a nonresponse.

Interview procedures for substitutes were the same as those for original cases. If a substitute residence refused to participate or could not be contacted, no additional substitute was generated for the original case.

Beginning with replicate six, it was decided that substitutes would not be selected in the final three days of a replicate because in earlier replicates such cases did not appear to have a realistic chance of being contacted and interviewed. Because of an error made in implementing this modification, no substitutes were selected in replicates six and seven. Therefore, the analyses cited in this report were based on ten replicates instead of twelve.

Four analyses of the substitution procedure used in this study were made. These analyses are described and the results are given in Section 4. Some conclusions and recommendations are given in Section 5.

4. Project Analyses and Results

Four analysis tasks were carried out in this investigation. These tasks, which are listed below, are discussed in subsections 4.1-4.4.

(1) Evaluation of the General Effectiveness of the Substitution Procedure.

This analysis included the calculation of the proportion of original cases that provided responses after being targeted for substitutes, the calculation of the proportion of targeted cases for which a substitute was contacted, and a comparison of the response rates of substitutes and of the original sample.

(2) Costs for Substitutes

Exact costs for substitution were not available from this study. However, several items closely related to costs were derived, including additional numbers of phone numbers, phone calls, interviews, and minutes associated with generating, pursuing, and interviewing substitutes.

(3) Comparison of Substitutes and Original Selections

The characteristics of 150 "late respondents" were compared to those of their substitutes. Comparisons were made for eight demographic and five health characteristics.

(4) Comparison of Variance Estimates Based on Substitution with those Based on Weight Adjustments.

This analysis consisted of a comparison of the two variance estimates for the estimated mean for each of five health characteristics.

4.1 Evaluation of the General Effectiveness of the Substitution Procedure

A total of 668 original sample units met the criteria for generating a substitute. Of these, 216 (32.3%) were eventually interviewed during followup. These 216 units are referred to as late cooperators or late respondents. Only 618 of the 668 substitutes needed were actually selected. Fifty substitute units were not selected because the need for them was not known until replicate closeout had arrived or, in later replicates, until the final three days before closeout. Of the 618 substitutes, 75 were never contacted. Of the 543 that were contacted, 435 were interviewed. For 150 of the 435 interviewed substitutes, interviews were also obtained from the original sample unit. These 150 comparative pairs of interviews formed the basis of the analysis discussed in Section 4.3. These counts are given in Table 1.

The response rate for substitutes was 74.0% as compared to 78.9% for the original sample.[3] The difference of 4.9% is because less time was generally available for contacting substitutes.

Regarding an evaluation of the rules for selecting subsitutes, an unclear picture is presented. Since 32.3% of the sample units targeted for substitutes were eventually interviewed, that substitutes may have been generated too early. But since 50 substitutes (7.5%) were never selected and since an additional 75 substitutes (11.2%) were never contacted, delaying the generation of subsitutes may not be wise.

Table 1. Breakdown of the Basic Counts for Substitution

Sample units targeted for substitution	668
Substitutes:	
Selected	618
Not selected	50
Selected substitutes:	
Contacted	543
Not contacted	75
Contacted substitutes:	
Interviewed	435
Partially interviewed	12
Refused	84
Other noninterviews	12
Interviewed substitutes:	
Original unit also interviewed	150
Original unit not interviewed	285

4.2 Costs for Substitutes

The exact costs incurred due to substitution were not available from this study. However, several items related to cost were derived in order to learn how much time and effort was expended in pursuing and interviewing substitutes. PSU averages were computed using the 208 PSU's that were selected from the 10 replicates used to study substitution. Table 2 summarizes these results.

The time spent on substitute cases can be viewed in terms of the equivalent number of original sample cases. Table 2 shows that an average of 125.16 minutes of on-line telephone time was used to pursue substitutes. For the original sample, the average amount of on-line telephone time spent per case was 45.90 minutes. Hence, the time spent per PSU on substitutes was equivalent to the time spent on approximately 2.73 (i.e., 125.16/45.9) original sample units. This average is slightly less than the average number of substitutes selected per PSU (2.73 vs. 2.97) because more time was generally available to pursue original cases than to pursue substitutes.

An important way of interpreting this data is that the survey funds used for the substitution procedure could have been used instead to increase the survey sample size by 3 units per PSU. This interpretation is critical for the variance comparisons presented in Section 4.4.

Table 2. Data on the Use of Substitutes

Item	Total from 10 replicates	Average per PSU
Number of times a substitute was supposed to have been generated	668	3.21
Number of substitutes actually selected	618	2.97
Number of additional phone numbers generated due to substitution (including ineligible cases)	1063	5.11
Number of additional phone calls made	3589	17.26
Number of additional interviews obtained	435	2.09
Number of minutes of on-line telephone time due to substitution	26033	125.16

4.3 Comparison of Substitutes and Original Selections

As indicated in Section 4.1, there were 150 matched cases for which interviews were obtained from both the original sample household and its substitute. This provided an opportunity to compare a population of late respondents with one of substitute respondents. Although this is not the same as the ideal comparison between all nonrespondents and their substitutes, this comparison is still useful because the late respondents would have been nonrespondents if follow-up attempts had been less extensive.

For the 150 pairs of original and substitute cases, a comparative analysis was carried out for eight demographic and five health characteristics. For four of the demographic characteristics and for all five of the health characteristics, a standard large-sample normal test was performed to see if the sample household means for the late respondents were significantly different from those for the substitutes. The nine characteristics included in this analysis are listed in Table 3 along with the sample means, the estimated standard error of the difference between these means, and the Z-score.[4]

Since the other four demographic characteristics are not quantitative, a comparison of means could not be made. Instead, a chi-square test was used for each of these characteristics to test the homogeneity of the original and substitute distributions. The characteristics are listed in Table 4 along with the computed chi-square statistic and the chi-square critical values for the 10% level of significance.

For both the comparisons of means and distributions, simple random sampling was assumed. Even though the full sample was selected in clusters of 12 units, the 150 pairs of late respondents and their substitutes are much less clustered than the full sample. Of the 102 clusters that contain at least one pair, 66 clusters contain exactly 1, 28 clusters contain 2 pairs, and 8 clusters contain 3 or more pairs. Therefore, the assumption of simple random sampling should not cause serious problems in this analysis although the standard errors of differences may be slightly underestimated.

Table 3 shows that a significant difference between the means at the 5% level existed for only two of the nine variables: age of reference person and average age of household members. In both cases the mean age of the substitutes was significantly higher than the mean age for the original cases. This implies that the ages of the persons in substitute households are generally higher than the ages of the person in the late respondent households. This is not surprising since the difficult-to-reach original sample households probably contain more younger and more mobile persons than the substitute households. Although no significant differences were observed between means for health characteristics, it is interesting that the average number of illness-related characteristics was always higher for the substitutes. This may also be due to the age differences.

For the four distribution comparisons summarized in Table 4, no significant differences were found at the 5% level. However, for sex of reference person the two distributions differed significantly at the 10% level, providing some evidence that late respondents and substitutes differ in this characteristic. This difference arose because the percent of female reference persons in the original sample (32) was significantly less than in the substitute sample (42). This suggests that substitute households contain disproportionately more female reference persons than do the late responding original households. A possible reason for this is that a higher proportion of men are in the labor force and are harder to contact than women.

4.4 Comparisons of Variance Estimates Based On Substitution with Those Based on Weight Adjustments

For each of the five health characteristics, a comparison was made between the variance estimate of the estimated mean based on the original sample plus substitutes and the variance estimate based on an equal-cost sample that used weight

Table 3. Comparisons of Means

Demographic Characteristics	Mean (Originals)	Mean (Substitutes)	Estimated Standard Error of Difference	Z-Score
Household Income	28,109	26,302	2,682	.67
Age (Reference person)	39.84	46.40	1.75	-3.75
Average age of household member	33.87	40.35	2.00	-3.24
Household size	2.39	2.44	.15	-.33
Health Characteristics (Number of)				
Hospital Stays in the Last Year	.105	.138	.034	-.97
Illness Bed Days in the Last Year	3.168	3.601	1.060	-.41
Doctor Visits in the Last Year	2.766	2.810	.466	-.09
Doctor Visits in the Last 2 Weeks	.201	.249	.057	-.84
Work Days Lost in the Last 2 Weeks	.086	.231	.101	-1.44

Table 4. Distribution Comparisons

Characteristic (of Reference Person)	Computed Chi-square Statistic	Ninetieth Percentile of Chi-square Distribution
Marital Status	5.21	7.78
Sex	3.22	2.72
Race	.31	4.61
Education	.76	7.78

adjustments, rather than substitutes, to account for nonresponse. It was demonstrated in Section 4.2 that if substitution were not used, three more telephone residences could have been selected per PSU with only a slight increase in costs. Therefore, the weight-adjustment sample that was taken to be equal in cost to the substitution-based sample was one consisting of the original sample of 12 residential units, plus three additional residential units per PSU.[5]

To obtain the additional cases, first the response rate for each PSU was calculated based on the original sample of 12 residences. This rate was multiplied by 3 to obtain the "expected number" of additional interviews if 15 residences had been selected initially. The expected number was rounded to the nearest integer to determine the number of additional interviews to add to the PSU. A constraint was included in this procedure so that the overall response rate for the augmented sample would equal that for the original sample.

The primary source of additional interviews were substitutes that had been interviewed for the PSU. For 107 of the 208 PSUs included in the analysis, there were enough substitute interviews available to provide the pseudo interviews needed needed. For each of the remaining 101 PSUs, one or more pseudo interviews were provided, as needed, by selecting cases randomly from the completed interviews obtained from the original selections. That is, entire interviews were "hot decked" (or replicated) to obtain the required number of additional interviews to complete the weight-adjustment sample. Three hot deck cases were needed for 16 of the 101 PSUs. For the other 85 PSUs, either one or two hot deck interviews were selected.

The weight-adjustment classes were the same as those used to generate substitutes-i.e., the individual PSUs. It was assumed that, by using the same classes for both procedures, the nonresponse bias for the substitution-based estimator of the mean would be about the same as that for the weight-adjustment-based estimator.[6] In this case, the nonresponse weight adjustment, w_i, assigned to each respondent selected from the i-th PSU is

$$w_i = k_i/k_i' , \qquad (1)$$

where

k_i = the PSU sample size (i.e., 15),
k_i' = the total number of completed interviews, including pseudo interviews, for the i-th PSU.

Since completed interviews were not obtained for all substitute cases, the weight adjustment given in equation (1) also had to be used for the sub-

stitution-based estimator. In this case, $k_i = 12$ and k_i' = the number of completed interviews in the PSU, including substitutes.

To develop the variance estimation expression, some notation is needed. First, the weighted sum, x_i', for a characteristic, X, for the i-th PSU is equal to

$$x_i' = w_i \sum_{j=1}^{k_i'} x_{ij} \quad ,$$

where x_{ij} = the sum of the values of X for all persons in the j-th respondent household in the i-th PSU. Similarly, the sum, n_i', of the weights for the i-th PSU is equal to

$$n_i' = w_i \sum_{j=1}^{k_i'} n_{ij} \quad ,$$

where n_{ij} = the number of persons in the j-th respondent household in the i-th PSU.

The population mean was estimated as follows:

$$\bar{x} = x'/n' \quad ,$$

where

$$x' = \sum_{i=1}^{208} x_i' \quad ,$$

$$n' = \sum_{i=1}^{208} n_i' \quad .$$

The variance estimate of the population mean, $\bar{x}$, was computed using the standard Taylor Series approximation to the variance of a ratio:

$$\hat{\sigma}^2_{\bar{x}} \doteq \bar{x}^2 \left(\frac{\hat{\sigma}^2_{x'}}{(x')^2} + \frac{\hat{\sigma}^2_{n'}}{(n')^2} - 2 \frac{\hat{\sigma}_{x'n'}}{x'n'} \right) \quad . \qquad (2)$$

All terms in equation (2) have been defined except the two variance estimates, $\hat{\sigma}^2_{x'}$ and $\hat{\sigma}^2_{n'}$, and the covariance estimate, $\hat{\sigma}_{x'n'}$. Each of these three estimates was derived using an ultimate cluster variance estimate. For example,

$$\hat{\sigma}^2_{x'} = \frac{208}{207} \sum_{i=1}^{208} (x_i' - x'/208)^2 \quad . \qquad (3)$$

The other variance estimate and the covariance estimate were computed in an analogous way.

For both the substitution-based and weight-adjustment-based estimators, the variances of the estimated means for all five health characteristics were estimated using equation (2). The ten variance estimates, along with the estimated means, are given in Table 5. The variance estimate for the substitution-based estimator was less than that for the weight-adjustment-based estimator for all five characteristics. Therefore, substitution appears to provide slightly lower variances than a PSU-by-PSU weight adjustment procedure.

However, further investigation has raised some doubts about the variance analysis summarized in Table 5. The method used to generate "pseudo cases" for the equal cost weight-adjustment sample may have provided misleading results. Specifically, the use of substitutes for additional interviews for the weight-adjustment sample may have introduced a component of variance that should not have been there. Consequently, the variance estimates given in Table 5 for the weight-adjustment procedure may be too high.

Table 5. Variance Estimates for the Substitution-Based and Weight-Adjustment-Based Estimates.

Health Characteristics (Number of)	Substitution		Weight Adjustment	
	Est. Mean	Est. Variance	Est. Mean	Est. Variance
Hospital Stays (Last Year)	.148	.000068	.152	.000077
Illness Bed Days (Last Year)	4.484	.107	4.553	.143
Doctor Visits (Last Year)	3.338	.0184	3.383	.0206
Doctor Visits (Last 2 Weeks)	.248	.00018	.247	.00020
Work Days Lost (Last 2 Weeks)	.247	.00074	.255	.00095

5. Conclusions and Recommendations

The general success of a substitution procedure will depend heavily on the rules used to initiate substitutes and on the call scheduling applied to substitutes. The substitution procedure used in this study was chosen primarily on an intuitive basis without much preliminary investigation. It turned out that this procedure was not particularly successful. A high portion (32%) of the cases targeted for substitution were eventually interviewed, which represents unnecessary expenditures. Perhaps there were certain types of cases for which substitutes were generated too early. Also, for 7.5% of the targeted cases, substitutes were never generated. Finally, the response rate for the substitutes that were generated was about 5% lower than for the original sample. The data collection period might have to be increased or the call scheduling modified to improve the generation rate and response rate for substitutes. Furthermore, consideration should be given to the possibility of generating additional substitutes for a case when the first substitute turns out to be a nonrespondent.

The comparison of hard-to-interview original sample cases and their substitutes, discussed in Section 4.3, addresses the potential for nonresponse bias in the use of sustitution to account for unit nonresponse. The reference persons in the substitute respondent households were older, had a higher percent female, and indicated a tendency to report higher numbers of illness-related activities than did their hard-to-interview counterparts. These differences indicate that there is the potential for biases in the

survey estimates due to the use of substitutes.

How would such biases compare to those associated with nonresponse weight adjustments in the case where adjustment classes are taken to be the same as the substitution classes (i.e., the individual PSUs)? In designing this research it was assumed that the biases associated with these two procedures would be about the same since substitutes are additional randomly selected units from the same PSU and weight adjustments impute the "average" characteristics of the respondents in the PSU to the nonrespondents in the PSU. However, since less time is generally available to pursue substitutes than original sample cases, substitute respondents must generally be "early cooperators." Consequently, there may be a bias component associated with the use of substitution that may not exist for the corresponding weight adjustment procedure. To minimize this differential effect, the rules for initiating substitutes, the interview period, and call scheduling procedure should be designed in such a way that adequate time will be available to pursue substitutes. The response rate for substitutes would provide an indication of whether there was adequate time to pursue them. If it were about the same as the response rate for the original sample, then there probably was adequate time provided to pursue substitutes.

With regard to variance estimation, discussed in Section 4.4, the comparison between substitution and weight-adjustment is unclear, due to the method used to generate three additional cases for the equal-cost weight-adjustment sample. Because of the potential of an "early cooperator" bias associated with substitution, the use of substitutes for pseudo interviews for the weight-adjustment sample may have added an erroneous component to the variance estimate based on the weight-adjustment procedure. Consequently, the relative sizes of the variance estimates for the two methods are uncertain.

To investigate the variance comparison further, within-cluster variances were computed for the substitution-based sample and for the weight-adjustment based sample, excluding pseudo respondents. These within-cluster variances, which were calculated using a standard formula from ANOVA methods, were computed for all five health characteristics. Also, the simple variance among the nonresponse weight-adjustment factors was calculated for both the substitution-based and weight-adjustment-based samples, again ignoring any pseudo cases. It turned out that the within-cluster variances were about the same for the two samples, as expected. However, the variance among the weight adjustment factors was only slightly higher for the weight adjustment sample than for the substitution sample. This was an unexpected result; it was assumed that the nonresponse weight adjustment factors would vary considerably more for the weight-adjustment sample, giving the substitution-based sample an advantage with respect to variances.

This surprising result regarding the variances of the nonresponse weight-adjustment factors, coupled with the result that the within-cluster variances are about the same for the two samples, suggests that the survey variances associated with substitution and weight adjustment may be about the same. Therefore, considering the potential for early cooperator bias associated with substitution, it appears that a substitution procedure for nonresponse should not be used for an RDD survey unless it can be designed so that the potential for early cooperator bias can be virtually eliminated.

More research is needed in this area, especially in terms of the bias and variance associated with substitution procedures, relative to nonresponse weight-adjustment procedures.

REFERENCES

Tegels, Robert and Chapman, David W. (1984). "Sampling Specifications for the National Health Interview Survey-Random Digit Dialing (NHIS-RDD) Feasibility Study," internal U.S. Bureau of the Census memorandum, January.

Waksberg, Joseph (1978). "Sampling Methods for Random Digit Dialing," JASA, 73, pp. 40-46, March.

FOOTNOTES

1 CATI is an acronym for computer assisted telephone interviewing.

2 The replicates were independent except that they were selected without replacement.

3 For deriving these response rates, it was assumed that a portion of the noncontacted cases were residential.

4 The standard error of the difference of means was estimated based on the 150 observed differences between late responding originals and their substitutes. The Z-score is simply the difference between means divided by the estimated standard error of the difference.

5 An equal-cost weight-adjustment sample could have been defined by retaining the fixed PSU sample size of 12 residences, but increasing the number of PSUs. However, it seems appropriate that the equal cost substitution-based and weight-adjustment samples being compared both contain the same number of PSUs per replicate. If additional PSUs are considered for the weight-adjustment sample, they should also be considered for the equal-cost substitution-based sample.

6 Even with equality of the substitution and weight adjustment classes, there is apparently an "early cooperator" bias associated with the substitution-based estimator but not with the weight-adjustment-based estimator. This is discussed in Section 5.