A COMPARISON OF FULL-PROBABILITY AND PROBABILITY-WITH-QUOTAS SAMPLING TECHNIQUES IN THE GENERAL SOCIAL SURVEY

GSS TECHNICAL REPORT NO. 5

by

C. BRUCE STEPHENSON

March, 1978

This research was supported by
National Science Foundation
Grant GS 31082X

The 1975 and 1976 General Social Surveys, conducted by the National Opinion Research Center (NORC), constitute among other things a unique experiment in sample design. Half of each survey used a multistage probability sample in which the choice of respondent was predetermined at every stage. The other half used a modified probability design with quotas at the block level, in which the interviewer canvassed a specific block seeking persons who met certain demographic criteria, and interviewed these persons until the quotas for that neighborhood were filled. For the sake of brevity, we will call these sample designs 'prob' and 'quota' respectively. They are perhaps the two most widely used sample designs in large-scale sample surveys, and the 1975-6 General Social Surveys (GSS) provide data for empirical comparisons of the two.

This report presents an exploratory analysis of the split-sample experiment. It is intended as a guide for people who are not primarily interested in sampling theory, but are concerned about the possible effects of sample design. We shall be trying to answer the questions of which variables have different response patterns in the two subsamples, and why they behave this way. The analysis is based on hundreds of statistical comparisons, which of necessity cannot be described in the following pages in more than the most general terms. In any case, significance tests alone are expected to be wrong five percent of the time. The main conclusions are based on both the empirical evidence and the existence of plausible explanations.

We will start with a brief description of the two sampling techniques. For a more detailed explanation, see Appendix A in any of the GSS codebooks. Some methodological considerations relating to the comparison will then be briefly discussed. Finally, we will present the results of the analysis,

subdivided into substantive and statistical results. The statistical section argues essentially that, aside from the variables discussed in the substantive section, the two sampling techniques produce pretty much identical results.

Sampling Designs

Both samples are based on the NORC National Probability Sample, a stratified multistage area probability sample of clusters of households in the continental United States. The Primary Sampling Units (PSUs) are either Standard Metropolitan Statistical Areas, as defined by the Bureau of the Census, or non-metropolitan counties. Secondary sampling units, or 'segments', are smaller areas chosen within each PSU. A third-stage unit is chosen within each segment as the sample is drawn for a given survey. Sampling of respondents, according to either method discussed below, takes place within the third-stage units. A typical GSS sample is drawn from approximately 100 PSUs, with 3 segments per PSU, and an average of 5 cases per segment, for a total of about 1,500 cases. Within any PSU, two segments were sampled with one technique and the third with the other. The exact number of respondents was 1,490 in 1975 (735 prob and 755 quota), 1,499 in 1976 (744 prob and 755 quota).

The Prob Subsample

In the subsample which we are designating prob, selection of households within the third-stage sampling units was done randomly from a complete listing of all of the households in the sampling unit. Interviewers went to the selected households and completed a four-page screener interview with a household member, obtaining information about all members of the household. This information enabled the interviewer to use a computer-generated selection table to choose the proper respondent. If necessary, an appointment was made to interview the selected person. Decisions at every stage of the selection process were thus independent of the availability of potential respondents.

The Quota Sample

In the quota subsample, a canvassing procedure determined which persons within the third-stage units were interviewed. Interviewers were given quotas based on sex, age, and employment status; they proceeded around a specified block from a specified starting point, seeking eligible persons within dwelling units and interviewing them as they were found. The quotas assigned a particular segment were calculated from 1970 census data for the area. Clearly, this method involved somewhat less control over the choice of respondent; however, the quota criteria were designed to produce representative numbers of certain hard-to-find population groups (young males, employed females).

Methodological Considerations

Before describing the analysis, we should note a couple of points which affect the difficulty of the task.

First, we are fortunate in that a large majority of the survey variables were present both years. We thus have two more-or-less independent samples, within each of which we can compare the prob and quota subsamples. (More-or-less independent because the PSUs were the same for the two surveys, while the second- and third-stage units were, in general, different.) The partial replication was exploited thoroughly, if not systematically, and was invaluable as a guide to which apparent subsample differences merited closer study.

Secondly, the problem is considerably complicated because we are dealing with cluster samples in which all of the cases from a given segment were selected with the same sampling method. Variables which are homogeneous within geographic (or any other) clusters can show large differences across any arbitrary division of the clusters, and in particular, across the arbitrary division of the clusters into prob and quota. To choose an artificial but clear-cut example, there are several size- and type-of-place codes on the surveys. These variables are completely defined by the segment from which a case is drawn; they are attributes of the segment, rather than of the individual case. If it made sense to ask whether quota sampling yielded more suburbanites that prob sampling[1], then any significance test would have to be evaluated with 300 'cases', not 1,500, since only 300 observations had been made. This complication can be dealt with by estimating design effects; details are discussed in the Appendix.

Finally, only single-variable effects (that is, subsample differences in the univariate frequency distributions) were considered. This was a practical limitation: all variables on the two surveys could be, and were, examined; but the number of pairs of variables made a complete examination of bivariate effects impossible. Users who suspect that any particular correlation(s) are affected by the difference between prob and quota sampling techniques are invited to analyze the data and find out whether their suspicions are justified.

Analysis

Two types of analysis were performed. First, all variables on each survey were crosstabulated with sample type, and the chi-square statistic was calculated to test the hypothesis that the observed distribution of the variable is independent of the sampling technique used to measure it. Variables with a large number of categories were collapsed into a suitable number for crosstabulation: age, for example, was collapsed into six categories, and the Census occupation codes were collapsed into seven major groups. This analysis has the merit of requiring no assumptions about measurement level. In addition, variables which are measured at or near the interval level (five-point scales were considered sufficiently detailed) were tested for a difference of means between subsamples. We shall discuss the crosstabulations first.

All variables for which the subsample-difference chi-square was significant at the .05 level were examined further. Results for the two years were compared to see if the apparent difference persisted, significant or not. The design effect was estimated to correct for the effect of clustering on significance estimates. Unfortunately, the statistics necessary to make the correction for clustering assume variables measured at the interval level. Variables for which the correction was necessary had to be forced into some form permitting the calculation of interval statistics; in many cases, a dichotomy. There was often no graceful way to dichotomize: the collapse was necessary precisely because subsample differences had been found, but the most natural ways of dichotomizing sometimes concealed these differences. As a conservative strategy, dichotomies were created so as to maximize the difference between subsamples; in

no case did the creation of a dichotomy eliminate a previously significant difference.

Design effects were estimated and used to correct significance estimates, as explained in the Appendix. A number of the corrected chi-squares remained significant. Some of these we believe to represent real differences between the sampling methods; these are discussed in the section on substantive results. The remainder we shall ascribe to chance in the section on statistical results.

The tests for differences in means were straightforward. Significant t values were divided by the square root of their design effect estimates, and their significance was then reevaluated. No new substantive discoveries were made. Specifics will be mentioned as appropriate in the two sections on results.

Substantive Results

There are two reasons why the response pattern for a variable might differ because of the sampling technique. Suppose that the probability that a person is selected as a respondent is related to some variable in a way which is different for each sampling technique. Then persons who score, for example, high on the variable will be over- or undersampled at different rates by the two techniques, so that high scores on the variable will be more likely in one sample type than the other.

The other possibility is analogous. If a person's probability of agreeing to the interview is related to some variable in different ways for the two samples, certain values of the variable will be more represented in one sample than the other. This is essentially an interaction between the variable, sampling technique, and response rate, just as the other reason involved an interaction between the variable, sampling technique, and probability of selection. The specific variables we have found which differ between the samples will clarify these abstract considerations.

1. Probability-of-Selection Interactions

In discussing probabilities of selection, it will be useful to focus on the household, since households are selected in the prob sample by a random process: nothing affects the probability that a household will be selected. In the quota sample, a household will be selected if (1) an interviewer, following the preassigned canvassing pattern, calls at the household; and (2) there is at least one person present who satisfies one of the quota categories. The chance of an interviewer calling is pretty much independent of household characteristics. The chance of

a suitable person being at home, however, is clearly related to the number of adults in the household: large households are more likely to include such a person. Large households are therefore more likely to be represented in a quota sample than in a prob sample. The relevant variable in the GSS is the number of adults in the household: over the two years, the mean number of adults per household was 2.22 in the quota subsamples and only 1.98 in the prob subsamples.

It should be noted that the unit of analysis in most survey research is the individual respondent, not the household. Rephrased in terms of individuals, the above discussion sounds somewhat different, although the bottom line is still the same. In the prob samples, each person within a selected household has the same probability of being the chosen respondent: 1/n, if there are n eligible adults. The individual's probability of selection is thus inversely proportional to the number of adults in the household. Persons from large households are also underrepresented in the quota samples (because, roughly, only one person per household can be interviewed regardless of how many satisfy the quotas), but the underrepresentation is less than in the prob samples.[2] The net result, again, is that more persons from large households will appear in quota samples than in prob samples.

This interaction between probability of selection, sampling technique, and number of adults per household also causes differences in related variables. The strongest cases are number of persons per household and number of persons earning money in the household; each year both of these variables reflect the greater household size sampled by the quota technique. Household size is also related to marital status, principally because the adult in single-adult households is rarely married. The prob

sample, by selecting more single-adult households, gets fewer married persons. (This difference was not significant either year, but was quite possibly real. The proportion married in the GSS dropped several percent from the all-quota 1974 survey to the all-prob 1977 survey; some of this difference may be secular change, however.) Large households are also more likely to be Catholic than small households, and each year the prob subsample had fewer Catholics than did the quota. Again, the differences were not significant but may be real.

One natural reaction to the fact that survey samples are biased against persons from large households is to weight the surveys in inverse proportion to the respondent's probability of selection, although weighting reduces the efficiency of the sample and is often found to make so little difference as to be unnecessary.[3] We have discussed such weighting elsewhere, arguing that if weights are applied, they should be applied to both types of sample, and deriving weights which compensate for differences in probability of selection due to household size.[4] As a check on the effects of such weighting, all of the subsample comparisons made for the present research were repeated with weights to compensate for the different probabilities of selection. The weights adjusted the household-size variables to rough equality between subsamples (not surprising, since the weights were computed to yield the proper distribution of adults per household); they eliminated the differences in proportion married and reduced the differences in proportion Catholic. Aside from this, weighting appears to have no effect. We shall therefore continue to discuss the unweighted comparisons.

2. Response-Rate Interactions

Normally, a characteristic which makes people uncooperative will be underrepresented in a sample survey. People who do not want to be interviewed will often refuse to be interviewed. In most cases, quota sampling will underrepresent grouchy people more seriously than prob sampling, because the field staff will put considerable effort into convincing a selected respondent to participate in a prob sample, while an interviewer for a quota sample can simply go next door. The GSS includes an interviewer rating of the respondent's cooperation, with four categories ranging from "friendly and interested" to "hostile." Each year, the proportion rated "friendly and interested" was about 4 percent higher in the quota sample than in the prob; these differences were not significant but are plausible. Of course, the interviewers generally had to go to somewhat more trouble to complete prob interviews than quota interviews, and they may have been simply expressing their irritation in this rating.

Grouchiness is not the only thing affecting response rate; some people are simply unavailable most of the time. Each sampling technique attempts to compensate for respondent unavailability, either by making appointments to call back at the respondent's convenience (in prob samples) or by canvassing only during hours when people are likely to be at home (in quota samples). Of course, the quotas themselves are designed to compensate for known differences in respondent availability. Let us consider them in turn.

Men are more difficult to find than women. Sex is therefore one of the variables controlled by the quotas. The prob sample selects people independently of their sex; however, both the 1975 and 1976 prob samples were about 58 percent women (compared to 52-53 percent in the quota samples).

Evidently, more men than women were refusing to be interviewed in the prob samples. In the 1977 GSS, which was all prob, the proportion of women had dropped to 55 percent, so perhaps the high proportions found in the 1975-76 prob subsamples were accidental. Nevertheless, it seems safe to assume that the quota controls can artificially produce a more accurate sex distribution than the random prob procedures, especially if the refusal rate really is higher among men.

People with jobs, particularly full-time jobs, are also difficult to find and interview. Indeed, the prob subsample includes more people with full-time jobs each year, although the difference is significant only in 1975. Since employment status is a quota control for women, but not for men, we made this comparison separately for the sexes. Among women there was no difference between the subsamples; among men, the differences are rather large. Combining the two years, 66 percent of the men in the prob subsamples but only 52 percent of the men in the quota subsamples were working full-time. The lack of a quota control for employment status among men clearly results in the more readily available men--those who are not working full-time--being interviewed.

Curiously, the other control which is applied to men in the quota sample, age, shows no similar effect. There are no age differences between the subsamples for either men or women. This presumably means that age quotas are unnecessary for women, and that the age quota for men is successful in securing the desired proportion of young men.

People who are grouchy, male, or employed are difficult to contact; as we have seen, this implies that they will be slightly underrepresented

in the prob samples, and more seriously underrepresented in the quota samples--unless the quota controls determine their representation. Characteristics which make people hard to contact and which are common to entire neighborhoods will be underrepresented, because of non-response, exactly as before in the prob samples, but will not be underrepresented in the quota samples. Quota interviewers will get their allotted number of interviews in a difficult neighborhood; prob interviewers, having no latitude to keep searching, may have to accept refusals. Something similar to this seems to happen in large central cities. The prob samples get fewer respondents than the quota samples in areas coded (1) on two size-of-place codes, NORCSIZE and SRCBELT. These areas are the larger central cities in the sample.[5] Almost inevitably, the differences are not significant; size-of-place codes are attached to neighborhoods, so their design effect is about 5. This type of neighborhood phenomenon is simply hard to measure. We can, however, look at cluster sizes achieved by the different sample techniques in different types of neighborhood. In the quota samples, average cluster size was about 4.9, regardless of neighborhood type. In the 1975, 1976, and 1977 prob samples, the average cluster size was 5.09. The average size of prob clusters coded (1) on NORCSIZE was only 4.13; the average size of prob clusters coded (1) on SRCBELT was only 3.69. A similar phenomenon seems to hold for New England clusters (average size 4.17).

We are suggesting that non-response in the prob samples occurs more often in big cities. Confirmation of this would require analysis of the location of non-respondents. We have not carried out this analysis. The differences discussed above are there each year, however, and non-response appears to be the most likely explanation.

Naturally, the slight underrepresentation of large cities in the prob samples affects certain characteristics which are concentrated in the cities. We may tentatively suggest three such characteristics. First, people who can name their ethnicity tend to live in large cities, and each year a greater proportion of the quota sample was able to name an ethnicity. Second, a greater proportion of the quota samples favored spending on problems of the cities, an attitude which is concentrated among people who live in cities. Finally, all religious groups except Protestants are concentrated in large cities to some extent; they are perhaps underrepresented slightly in prob samples.

Statistical Results

It is difficult to summarize the statistical analysis on which this paper is based, principally because of its bulk. The results presented in the previous section were extracted from hundreds of comparisons, which cannot be described here in any detail. We hope that most of the important differences between the two sample types have been covered, but there are undoubtedly others. Since GSS data are widely available (all of the analysis reported here except the estimation of clustering statistics can be made from the public GSS datasets), we shall present only a bare summary.

Table 1 gives the raw chi-square, design effect, corrected chi-square, degrees of freedom, and significance for all variables which had raw chi-squares significant at the .05 level. As mentioned earlier, these variables had to be dichotomized so that design effects could be estimated, unless they were already measured at or close to the interval level; raw chi-squares in the table are for the dichotomized variable in each case. Some of the variables which remain significant have been discussed in the section on substantive results (WRKSTAT, SEX, ADULTS). We believe that these variables, along with HOMPOP and EARNRS, differ between the subsamples. The PHONE difference is probably an artifact; more people reported having telephones in the 1975 prob subsample, but prob respondents were asked twice (once in the screener interview and once in the main interview) and a positive response either time was accepted. Subtracting these 12 comparisons (6 variables, 2 years) from the 467 cross-tabular comparisons originally made leaves 455 comparisons made on variables which are not believed to differ between subsamples.

Excluding differences in the variables mentioned above, twenty-three (or .051 of the total) corrected chi-squares are significant at the .05 level. Five (.011 of the total) are significant at the .01 level. Moreover, several of the significant differences involve the highly artificial dichotomies which were created to preserve subsample differences while permitting the calculation of design effects. These remaining variables, then, appear to be randomly distributed between the subsamples.

Turning to the interval-level tests for differences in means, we have in Table 2 the t-value, the square root of the design effect, and the corrected t-value and its significance for all variables where the uncorrected t was significant at the .05 level. Degrees of freedom are not given; with hundreds of cases, t is distributed normally. We acknowledge that ADULTS, HOMPOP, and EARNRS differ between the subsamples; removing these 6 comparisons from the 147 originally made, we are left with 141 comparisons. Twenty-two of these (.156) are significant at the .05 level.

It is quite clear that several of the significant differences between means cannot be real differences due to sampling technique: they change directions between 1975 and 1976. In fact, a great many of the interval-level variables in the GSS are measures of essentially the same thing. It is a fact that the 1975 prob sample ranked higher than the 1975 quota sample on almost all variables related to social, economic, or educational status. Fortunately, it is also a fact that the 1976 prob sample ranked lower than the 1976 quota sample on all of the same variables. We evidently have a neighborhood phenomenon: a majority of

high-status neighborhoods fell into the prob sample in 1975, the quota sample in 1976. The reversal is lucky, for without it we might have been left with all sorts of fears about education or income biases in one or the other type of sample. The size of the reversal is slightly embarrassing, since we have measured it so many times. If there are no status differences in the universe, why did so many turn up in our samples? This question necessitates a brief digression in defense of probability theory, which may profitably be skipped by all readers who are not bothered by the question.

As hinted above, the reason so many variables which do not really differ between subsamples showed significant differences these two years is just that the variables all measured the same thing. The significance tests are trying to lead us into a "Type I" error, as they will do 5 percent of the time, but many of the instances in Table 2 are the same "Type I" error.

The variables which seem particularly likely to be involved in this sort of thing are education of respondent, spouse, father, and mother; Hodge-Siegal-Rossi prestige of respondent's and spouse's jobs; family and respondent income; educational requirements and relationship to data of respondent's and spouse's jobs; and Temme prestige of respondent's and spouse's jobs. These are all, of course, different variables. They tend to be highly correlated, however, and highly clustered in neighborhoods. A complete principal-components analysis of the 14 variables was done each year to test the idea that a shift in "status" as measured by all 14 of them, was responsible for the observed differences. The complete principal-components solution has the characteristic that the 14 factors are merely a linear transformation of the 14 original

variables; no information has been lost or gained. The factor solutions for the 2 years were almost identical; each year the first factor accounted for 47-49 percent of the variance, with most loadings above .7. Each year this principal factor was highly clustered, with a design effect over 2.0. If we replace, each year, these 14 variables with the 14 equivalent but statistically independent factors, only 2 of the 28 subsample comparisons are significant, instead of the original 15. (The principal factor is 'significantly' higher in the prob subsample in 1975; the second factor, which involves residual variation in respondent's occupation as contrasted with the spouse's occupation, is 'significantly' higher in the 1976 quota sample. Neither difference was consistent across the 2 years.) Details of the factor analysis are readily reproducible from the public data.

Counting 2, instead of 15, significant differences on these 14 variables for the 2 years, we are left with 9 of 141, or .064, of our comparisons significant at the .05 level, and 2, or .014, significant at the .01 level.

To conclude this statistical summary, it appears possible that the differences noted as substantive results may be only real differences produced by sampling techniques in the univariate response patterns on the 1975-1976 General Social Surveys.

Conclusion

Of the substantive results discussed above, three are fairly important. First, selection procedures lead to different distributions of household size. The quota sample overrepresents large households, while both samples, especially the prob, underrepresent persons from large households. Second, the quota sample underrepresents men who are working full-time. Finally, the lower response rate achieved in large cities leads to their being underrepresented in the prob sample. All three of these differences affect related variables, but the indirect effects are small, typically on the order of a percent.

Significance testing is of limited usefulness when one analyzes hundreds of variables. The design effect estimates used are only approximations, but precise significance tests were not crucial to any of the findings mentioned above. The fact that only .051 of the null-hypothesis chi-squares, and .064 of the t-tests, were significant at the .05 level suggests that the estimates were not far off. The slight excess may mean that we underestimated design effects; it may mean that some real differences were missed; it may mean nothing at all.

The limitations of this research should be acknowledged. We have considered only univariate response patterns, and only the variables which were on the 1975 or 1976 General Social Surveys. We do not expect that multivariate effects will be large; it seems quite possible, however, that other variables exist which are affected by the type of sample with which one measures them. Finally, we have completely ignored the purely statistical question of whether it is possible to justify any significance tests at all in a sample which involves quotas. The

probability sample with quotas at the block level, as used in the GSS, appears to work quite well generally. Like the full-probability sample, it has peculiarities which researchers should recognize; we hope that the more important of these have been identified here.

APPENDIX:

Estimation of Design Effects

Estimation of design effects was of particular importance for the analysis reported here: hundreds of subsample comparisons were made, and reasonable estimates of the corrections to simple-random-sample (SRS) significance tests were necessary if the research was to make any progress at all. Exact statistics for complex sample designs are difficult and in some cases unknown; they were not attempted. We shall describe here the nature of the problem, and the way in which the necessary statistics were estimated. This should not be construed as an adequate or precise discussion of the statistics of complex samples,[6] but only as an explanation for the statistics used in the text and tables.

Simple-random-sample significance tests are inappropriate to survey samples such as those used on the GSS principally because of the clustering of cases into segments and PSUs. Clustering does not bias the estimates of such population parameters as the mean and the variance; however, it increases the standard errors of such parameters. The expected value of the mean computed from a cluster sample is still the population mean, but repeated sample means will fluctuate more than would the means of simple random samples. The result, broadly speaking, is that confidence intervals and significance tests made according to the usual simple-random-sample assumptions overestimate the precision or significance of sample statistics. The most common measure of this overestimation is the design effect for the variance of the mean, defined as the ratio of the variance of the sample mean to the variance it

would have in simple random samples of the same size. A related concept is the 'effective number of cases', which is the actual number of cases divided by the design effect. The effective N is the number of cases which a simple random sample would require to give an equally precise estimate of the population mean.

The design effect for a cluster sample can be estimated by performing an analysis of variance by PSU, and dividing the between-PSU mean square by the total mean square (which equals S^2, the population variance estimate). To simplify the demonstration, assume that the clusters formed by PSUs are of equal size. Let

x_{ij} = value of some variable X for case j in PSU i

n = number of cases per PSU

k = number of PSUs

$N = nk$ = total number of cases

$\bar{x}_i = \frac{1}{n} \sum_{j=1}^{n} x_{ij}$ = mean of X in PSU i

$\bar{\bar{x}} = \frac{1}{k} \sum_{i=1}^{k} \bar{x}_i = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n} x_{ij}$ = sample mean of X

Consider the k cluster means we have for the PSUs. From these observations we can estimate the variance of the 'population' of sample means of size n, just as any other population variance could be estimated from k observations:

$$\text{Var}(\bar{x}_i) = \frac{\sum_{i=1}^{k} \bar{x}_i^2 - \frac{\left(\sum_{i=1}^{k} \bar{x}_i\right)^2}{k}}{k - 1}$$

$$= \frac{\sum_{i=1}^{k} \left(\frac{\sum_{j=1}^{n} x_{ij}}{n} \right)^2 - \frac{\left(\sum_{i=1}^{k} \frac{\sum_{j=1}^{n} x_{ij}}{n} \right)^2}{k}}{k - 1}$$

$$= \frac{\sum_{i=1}^{k} \frac{\left(\sum_{j=1}^{n} x_{ij} \right)^2}{n^2} - \frac{\left(\sum_{i=1}^{k} \sum_{j=1}^{n} x_{ij} \right)^2}{n^2 k}}{k - 1}$$

$$= \frac{1}{n} \left[\frac{\sum_{i=1}^{k} \frac{\left(\sum_{j=1}^{n} x_{ij} \right)^2}{n} - \frac{\left(\sum_{i=1}^{k} \sum_{j=1}^{n} x_{ij} \right)^2}{nk}}{k - 1} \right]$$

The quantity in brackets is just the between-PSU mean square from the analysis of variance. Denote it by BPMS:

$$\text{Var} (\bar{x}_i) = \frac{1}{n} \text{ BPMS} \qquad (1)$$

The overall mean, $\bar{\bar{x}}$, is normally viewed as the mean of the entire sample of N = nk individual observations. We wish to estimate its variance; however, the individual observations are clustered. If they were not clustered, we could use the standard formula for the simple-random-sample mean:

$$\text{Var}_{srs} (\bar{\bar{x}}) = \frac{S^2}{N} \qquad (2)$$

In a cluster sample, however, we must view $\bar{\bar{x}}$ as the mean of a sample of k PSU means. Its variance is then estimated as

$$\text{Var}\left(\bar{\bar{x}}\right) = \frac{1}{k}\,\text{Var}\left(\bar{x}_i\right) = \frac{1}{k}\,\frac{1}{n}\,\text{BPMS} = \frac{1}{N}\,\text{BPMS} \quad (3)$$

Comparing (2) with (3), we see that the design effect is estimated by

$$\text{Design effect} = \frac{\frac{1}{N}\,\text{BPMS}}{\frac{1}{N}\,S^2} = \frac{\text{BPMS}}{S^2} \quad (4)$$

Of course, the estimated design effect for a variable will fluctuate from sample to sample. In particular, the between-PSU mean square is based on only k - 1 degrees of freedom (about 100 in the GSS), so standard errors based on the calculated BPMS will fluctuate more than would the naive simple-random-sample estimates. They are, however, more accurate measures of the precision of the cluster-sample mean.

Application of the calculated design effect to the tests for difference in means is simple. The variance of the mean equals the SRS estimate times the design effect, so the standard error equals the SRS estimate times the square root of the design effect, and the corrrected t value equals the SRS estimate divided by the square root of the design effect.

The chi-square calculated from a crosstabulation varies linearly with sample size: multiplying each cell by some number A multiplies the chi-square by A. Chi-squares were therefore corrected by dividing the raw chi-square by the design effect, which has the effect of reducing the sample size to the effective number of cases as defined above. This procedure is not strictly correct, since we have estimated the design effect for the variance of the mean, which is not necessarily

the same as the design effect for chi-square significance tests. Adequate statistical formulations are simply not available for significance tests other than between means. However, any differences between a sample's effective size for t-tests and for chi-squares are probably inconsequential compared to fluctuations in the design effect estimates. The correction seems to have worked well, since the proportion of Type I errors in the corrected chi-square tests turned out to be about what was expected.

One further point is worth mentioning. There is some evidence that where a division cuts through clusters, the design effect for the difference in means across that division is smaller than the simple design effect for the mean.[7] The division we have been considering, between the two subsamples, does cut through the PSUs, but does not cut segments: all segments are entirely within one or the other subsample. The analysis of variance revealed that most of the clustering is within segments. It therefore appears likely that overestimation of design effects due to the splitting of PSUs is not a serious problem.

FOOTNOTES

[1]This question doesn't make sense simply because cases are determined as suburban or not suburban by the segment in which they fall; prob and quota techniques are applied only within segments. It does make sense to ask whether suburban segments (or small-town segments, or segments in the South) are pre-allocated proportionally to the two subsamples, although even this question is irrelevant unless sampling technique affects other variables of interest. See also the discussion of differential response rates later in this paper.

[2]Biases in probability of selection due to household size are discussed at greater length in GSS Technical Report No. 3, cited below.

[3]Leslie Kish, Survey Sampling (New York: John Wiley & Sons, 1965), page 400.

[4]B. Stephenson, "Weighting the General Social Surveys for Bias Related to Household Size," GSS Technical Report No. 3 (NORC, 1978), photocopy.

[5]Code (1) on NORCSIZE represents central cities with a population of over 250,000. Code (1) on SRCBELT represents central cities of the twelve largest SMSAs in the country. For evidence that respondents in large cities are difficult to contact, see W.C. Dunkelberg and G.S. Day, "Nonresponse Bias and Callbacks in Sample Surveys," Journal of Marketing Research, X (May, 1973), pages 160-168.

[6]See Kish, especially Chapter 5, for a thorough discussion.

[7]See Leslie Kish and Martin Frankel, "Inference from Complex Samples," Journal of the Royal Statistical Society, series B, vol. 36, pages 1-37.

TABLE 1

VARIABLES FOR WHICH THE RAW SUBSAMPLE-DIFFERENCE CHI-SQUARE WAS SIGNIFICANT

Variable[a]	Year	Chi-square	DEFF	Corrected Chi-square	Degrees of Freedom	Corrected Significance
WRKSTAT	75	6.19	1.09	5.68	1	.05
INDUSTRY	75	14.30	1.06	13.49	1	.01
MARITAL	75	5.67	1.82	3.12	1	
AGEWED	75	4.86	1.26	3.86	1	.05
DIVORCE	75	5.59	1.73	3.23	1	
SPHRS	76	7.38	1.26	5.86	1	.05
SPPRES	75	4.38	1.19	3.68	1	
PAIND16	76	17.23	1.05	16.41	1	.01
AGE	76	8.93	1.12	7.97	1	.01
SPEDUC	75	10.32	1.43	7.22	1	.01
SPDEG	75	9.30	1.06	8.77	1	.01
SEX	75	5.74	.63	---	1	.05
RACE	75	6.80	3.45	1.97	1	
FAMILY16	76	4.60	1.03	4.47	1	.05
ETHNIC	76	12.39	2.00	6.20	1	.05
ETHNUM	76	7.20	2.46	2.93	1	
HOMPOP	75	23.27	1.65	14.10	8	
"	76	16.84	1.78	9.46	8	
ADULTS	75	26.54	1.48	17.93	3	.01
"	76	45.90	1.87	24.55	3	.01

[a]See GSS Codebooks for question wordings. All variables in the table appeared in both the 1975 and 1976 surveys except SPKATH, COLMIL, SPKHOMO, LIBHOMO, MEMSERV, MEMUNION, CHLDIDEL, and TVHOURS.

TABLE 1-Continued

Variable	Year	Chi-square	DEFF	Corrected Chi-square	Degrees of Freedom	Corrected Significance
EARNRS	76	15.29	1.87	8.18	4	
INCOME	76	24.15	2.25	10.73	3	.05
RINCOME	75	9.35	1.26	7.42	3	
WORDI	76	5.11	1.42	3.60	1	
NORCSIZE	75	33.28	4.95	6.72	8	
"	76	38.07	4.98	7.64	8	
SRCBELT	75	22.35	4.95	4.52	5	
"	76	27.25	4.98	5.47	5	
VOTE72	76	6.61	1.54	4.29	1	.05
PRES72	75	4.92	1.44	3.42	1	
POLVIEWS	75	13.75	1.31	10.50	6	
NATENVIR	76	4.95	1.95	2.54	1	
NATCITY	75	4.00	1.52	2.63	1	
"	76	8.72	1.66	5.52	1	.05
NATCRIME	76	5.38	1.36	3.96	1	.05
NATFARE	75	5.49	1.69	3.25	1	
SPKATH	76	11.72	2.15	5.45	1	.05
COLMIL	76	4.23	1.58	2.68	1	
SPKHOMO	76	4.36	2.10	2.08	1	
COLHOMO	76	10.39	2.29	4.54	1	.05
LIBHOMO	76	6.09	2.13	2.86	1	
GRASS	76	4.07	1.49	2.73	1	

TABLE 1-Continued

Variable	Year	Chi-square	DEFF	Corrected Chi-square	Degrees of Freedom	Corrected Significance
USINTL	75	5.99	1.45	4.13	1	.05
RACPUSH	76	7.22	2.35	3.07	1	
RACLIVE	75	4.24	3.80	1.12	1	
HEALTH	76	10.57	1.57	6.73	3	
FAIR	75	8.53	1.88	4.54	1	.05
CONARMY	76	7.49	1.33	5.63	2	
MANNERS	75	9.60	1.47	6.53	4	
CLEAN	75	9.70	1.19	8.15	3	.05
RESPONSI	75	11.65	1.64	7.10	4	
MEMSERV	75	5.56	1.00	5.56	1	.05
MEMUNION	75	4.14	1.77	2.34	1	
CLASS	75	13.83	2.41	5.74	1	.05
"	76	8.60	1.98	4.34	1	.05
UNEMP	75	4.55	1.44	3.16	1	
ABRAPE	76	4.36	1.45	3.01	1	
CHLDIDEL	75	12.59	1.39	9.06	5	
PORNOUT	75	4.79	1.27	3.77	1	
HITOK	75	4.46	1.65	2.70	1	
TVHOURS	75	25.15	1.41	17.84	9	.05
PHONE	75	12.97	1.79	7.25	1	.01
COMPREND	76	10.78	1.47	7.33	2	.05

TABLE 2

VARIABLES FOR WHICH THE UNCORRECTED DIFFERENCE OF MEANS WAS SIGNIFICANT

Variable[b]	Year	t[c]	$\sqrt{DEFF}$	Corrected t[c]	Corrected Significance
PRESTIGE	75	-2.80	1.13	-2.48	.05
SPPRES	75	-3.33	1.18	-2.82	.01
SIBS	75	3.09	1.26	2.45	.05
EDUC	75	-3.13	1.41	-2.22	.05
"	76	2.78	1.40	1.99	.05
PAEDUC	76	1.97	1.38	1.43	
MAEDUC	75	-2.95	1.52	-1.94	
SPEDUC	75	-3.95	1.30	-3.04	.01
HOMPOP	75	3.27	1.28	2.55	.05
"	76	3.36	1.33	2.53	.05
TEENS	76	2.08	1.07	1.94	
ADULTS	75	4.79	1.22	3.93	.01
"	76	6.77	1.24	5.46	.01
UNRELAT	76	2.34	1.15	2.03	.05
EARNRS	76	3.84	1.37	2.80	.01
INCOME	75	-1.98	1.65	-1.20	
"	76	4.31	1.63	2.64	.01
RINCOME	75	-3.04	1.10	-2.76	.01

[b]See GSS Codebooks for question wordings. All variables in the table appeared in both the 1975 and 1976 surveys except WORDSUM, CANADA, ISRAEL, SOCOMMUN, SOCFREND, and TVHOURS.

[c]Positive values of t indicate higher mean in the quota subsample; negative values indicate a higher mean in the prob subsample.

TABLE 2-Continued

Variable	Year	t	$\sqrt{DEFF}$	Corrected t	Corrected Significance
RINCOME	76	2.00	1.22	1.64	
WORDSUM	76	3.09	1.37	2.26	.05
DOTDATA	75	2.45	1.12	2.19	.05
"	76	-2.32	1.14	-2.04	.05
DOTGED	75	-3.15	1.22	-2.58	.01
"	76	2.68	1.22	2.20	.05
DOTPRES	75	-3.38	1.29	-2.62	.01
"	76	2.09	1.28	1.63	
SPDOTDAT	75	2.83	1.22	2.32	.05
SPDOTGED	75	-2.98	1.22	-2.44	.05
SPDOTSVP	75	-2.04	1.20	-1.70	
SPDOTPRE	75	-3.14	1.23	-2.55	.05
CANADA	75	2.28	1.04	2.19	.05
ISRAEL	75	2.00	1.05	1.90	
CLEAN	75	-2.05	1.09	-1.88	
AMICABLE	76	-2.43	1.09	-2.23	.05
RESPONSI	75	2.61	1.28	2.04	.05
SOCOMMUN	75	2.25	1.23	1.83	
SOCFREND	75	2.05	1.07	1.92	
TVHOURS	75	3.39	1.19	2.85	.01