SAMPLING BIAS IN STATEWIDE TELEPHONE SURVEYS OF HIGH SCHOOL DROPOUTS

Terry F. Buss, Youngstown State University

Much public opinion, marketing and advertising research depends on responses from telephone interviews. Telephone directories are typically used to produce a sample of potential respondents for these interviews and this may lead to possible problems in producing a sample of completed interviews which is representative of a target population. Problems in representativeness in turn may be exaggerated if special or hard to locate populations (e.g., Sudman, 1972) are being surveyed and if statewide or large area sampling is required (e.g., Glasser and Metzger, 1972).

The purpose of this paper is to discuss the possible sampling biases which may result in telephone directory sampling of the population of high school dropouts for the State of Ohio.

Sampling in Telephone Surveys

Sampling problems in telephone surveys may revolve around telephone directories and respondents. The use of telephone directories in statewide samples produces immense problems for sampling enterprises. First, many respondents may not have listed telephone numbers or may have no telephone at all (e.g., Brunner and Brunner, 1971; Perry, 1968-1969). In some areas, for example, unlisted numbers could include from 20 to 30 percent of the households having telephones (Glasser and Metzger, 1972). Unlisted numbers or no telephone may also be characteristic of certain socio-economic groups who may have even higher incidents of unlisted numbers. This appears to be especially true for blacks, the poor, single women, members of unions and so on (Klecka and Tuchfarber, 1974; Glasser and Metzger, 1975).

Second, telephone numbers which are not unlisted, but which have not been published in a current directory are also problematic. Even at the time of publication of a directory, many "new" telephone numbers may be excluded. This results from persons moving into an area or persons moving within an area. The percentage of new numbers may be in part a function of the age of the telephone directory. In statewide sampling, the problem is magnified because individual directories from across the state will be of varying publication dates so that bias from this source will not be consistent for all directories used.

Third, the problem of new numbers is counterbalanced by that of disconnected numbers or numbers which have been reassigned (e.g., Glasser and Metzger, 1972). In the former case, numbers are called but no answer or a recording is heard. In the latter, the person contacted may not be an appropriate respondent for the sample.

Fourth, statewide sampling produces some problems which arise because of the large number of directories which are required to produce a sample. One of these involves the independent telephone companies in a state. An initial problem is that independent directories are sometimes difficult to obtain (Glasser and Metzger, 1972). Once obtained, it is frequently the case that calls made to a community served by an independent must be routed through several operators before a number is reached. Once a given number is reached, one may discover that telephone exchanges have been changed without notice or that numbers simply are unreachable from another telephone system.

Another problem which is intensified, but not attributable to independent companies is that of multiple listings. Our study revealed that a single number could be listed in as many as three different telephone directories. This is most likely to occur in areas served in part by independents in rural areas which border urban areas: this results from one company listing adjacent area numbers as a courtesy. It may also occur in directories which overlap state boundaries.

An entirely different set of problems may arise with regard to respondents. Problems here arise from the inability of interviewers either to contact appropriate respondents or to complete interviews with respondents once contacted. A major problem involves the inability of an interviewer to contact anyone at a given telephone number. This refers to "no answer". Another problem is the respondent who initially refused to be interviewed (e.g., Dillman, et al., 1976). It may not be clear whether or not the refusal was a member of the target population. Even if a respondent is in the target population, he may terminate an interview in the middle thereby rendering his previous responses useless. Yet another problem is an ineligible person refusing to call a potential respondent to the telephone to be interviewed. And finally, depending on the target group, a good many ineligible respondents may be called before an eligible respondent is reached.

All of these factors above either lead to more extensive time/resource commitments to a project, or lack of predictability about project requirements in sampling, or potential bias in the resulting sample. They are probably characteristic of most telephone surveys, but are perhaps more exaggerated in statewide surveys of high school dropouts.

Socio-economic Characteristics of High School Dropouts

High school dropouts have several characteristics more or less in common which make them extremely difficult to locate or interview by telephone (Boggs, et al., 1978). In general, high school dropouts tend to be low socio-economic status individuals. They are highly mobile in the sense of changing residences frequently. Because they frequently find themselves in debt or the subject of observation by various governmental agencies (e.g., welfare, education, social services, etc.), they may be very suspicious of contact from persons outside their family or peer group.

These characteristics have the following implications for sampling from telephone directories and interviewing by telephone. Many cannot afford telephones. As a result, they will be automatically excluded from the sample. Those who have telephones frequently prefer unlisted numbers so that they cannot be "harassed" by debt collectors or government agencies. They

will also be excluded unless random digit dialing or some other technique is used to locate them (Glasser and Metzger, 1972). Those who have listed numbers will be difficult to locate because of their mobility. In one study, 60% of this group had changed addresses at least once during a two year period and most failed to leave any forwarding address (Boggs, et al., 1979).

Assuming that a high school dropout has been reached on the telephone, it may be difficult to complete an interview. This group is very suspicious of outsiders prying into their personal lives or even soliciting information from them. Many think that an interview over the telephone, for example, represents some devious way of locating them by debt collectors or by social agencies. Consequently, they are likely not to participate in an interview or are likely to terminate an interview if questioning seems threatening (Boggs, et al., 1979).

Methodology

The State of Ohio was initially stratified into 14 Standard Metropolitan Statistical Areas (SMSA's) and into non-SMSA areas identified by county (see Bureau of Census, 1972). The SMSA and non-SMSA areas were then grouped in nine individual strata so that each area contained at least one SMSA surrounded by a non-SMSA area. An additional stratum was composed entirely of non-SMSA areas. A total of 35 non-SMSA areas were included in the sample design.

A complete set of telephone directories for the state was assembled. The set was then grouped into the 10 areas above. All SMSA telephone directories were automatically included in the sampling process. Directories containing non-SMSA telephone numbers were organized into three categories (large, medium and small) by population size. A random sample of directories from each category was selected. The proportion of the population for the state as represented by each directory was calculated next and used to estimate the number of completed interviews which would be required from each area.

(Place Table 1 About Here.)

Once the directories and their proportional contribution was identified, a special computer program was written which could be used to produce a systematic random sample (Sudman, 1976) of locations of telephone numbers in each directory used in the sampling process. The program produced page numbers, column numbers (one through five), column locations (top or bottom). Researchers then went through each telephone directory and recorded the designated numbers. Approximately 15 telephone numbers were drawn for every expected completed interview yielding a total of 22,500 numbers for 1,500 interviews. Care was taken so that numbers which were listed in more than one directory were not included more than once in the sample. Business numbers were not included in the sample.

High school (and grade school) dropouts were identified in the questionnaire introduction by means of two filter questions with accompanying "probes" by interviewers. Respondents were asked: "What is the highest grade you have completed in school?" followed by "Are you still attending school?" and "In what year were you born?" These questions served to identify approximately 95% of the appropriate respondents. The 5% remainder were identified by cautious probing.

This rather extensive questioning process was viewed as necessary for two reasons. First, some respondents may not readily admit that they are school dropouts. Second, those who may admit that they are dropouts also might be offended at the initial asking of a question concerning whether or not they dropped out. Since so many telephone calls would be required to reach this group, it was also felt that embedding a question about dropping out further on in the questionnaire would waste valuable time in locating appropriate respondents.

Interviews with 1,503 respondents were completed between May 24 and July 5, 1977. Approximately 13 telephone numbers were called per completed interview. This yielded 19,539 (1,536 x 13) separate telephone numbers for use in the study. Since up to three callbacks were permitted per telephone number in order to produce a completed interview, approximately 39,936 telephone calls were made in the study.

Findings

Representativeness of Sample

Analysis began with a comparison of the demographic characteristics of the high school dropout sample with equivalent data for the 1970 Census (see Bureau of Census, 1972). Representativeness of the sample for purposes of this paper was determined on the basis of race, age and sex.[1]

Data reveal results which were not entirely unexpected. Females were greatly overrepresented in the sample as compared with the 1970 Census. This was the case regardless of age or race of respondents. Apparently, the problem of oversampling women in telephone surveys (Klecka and Tuchfarber, 1974) also occurs in the specialized population of high school dropouts. Women comprised 77.0% of the sample, while men comprised 23.0%. In this study, it was not possible to ask to speak to a male at the outset of the interview because of the elaborate series of filter questions which were required to identify dropouts and because of the substantial increase in interviewer time which would have been required. Care was taken, however, to call during evening hours when men would most likely be at home. Therefore, it appears that data based on sexual composition would have to be weighted heavily to account for the overrepresentation of women.

Older persons were also overrepresented in the sample (31.0% were over 65 years of age) regardless of sex or race. At the same time, younger persons were greatly underrepresented (only 7.2% were between the ages of 16 to 24 years). Again this is a common problem with telephone interviewing. Because the population of high school dropouts was small and difficult

to locate, no attempt to correct sample proportions for age was made during the initial stages of the interview. Again, time/cost factors would have been prohibitive. Samples of this kind, then, must be insighted for age of the respondent once completed.

Racial factors in the study were not as biased as might be expected. For blacks (14.4%) and Spanish-surnamed (1.2%) respondents, sample results were not too far deviant from 1970 Census data expectations. For whites (84.1%), however, younger males and to a somewhat lesser extent, younger females were greatly underrepresented.

Overall results, then, showed that the most numerous, high school dropouts to reach was young, white and male. This particular individual probably lives at home, but has no telephone of his own. He spends a good deal of time away from home. He is difficult to pin down for an interview by telephone at anytime. As a result, future surveys of this subpopulation must find ways to include larger numbers of respondents in the sample. Otherwise, extensive weighting of this group in the sample must occur.

When data was controlled for sampling area sexual, racial and age differences persisted in a fashion consistent with the overall pattern. Apparently, then, sampling biases resulting in these three dimensions were fairly consistent across various subdivisions of the State of Ohio.

Some Characteristic Results of Telephone Interviewing

Analysis next considered the relationship between sampling area and differences in completion rates, initial refusals, within interview withdrawals, non-respondent refusals, disconnected numbers, ineligible respondents and no answers. Table 2 reveals the results of this analysis.

(Place Table 2 About Here.)

Table 2 shows that overall, "no answers" when attempting to complete interviews are the most important factors in depleting telephone numbers in use in the sample. In this study, 43.7% of the numbers called yielded "no answer" by any individual. A great deal of variation around this grand total was observed across SMSA's with a low of 26.8% to a high of 59.6%. The greatest difference in "no answers" was between SMSA (38.9%) and non-SMSA's (59.3%). Results, therefore, suggested that it may not be possible to accurately predict or allow for "no answers" in deciding how many telephone numbers to include in a sample; but that rural areas in relative terms must be greatly oversampled as compared to more urban areas.

Disconnected telephone numbers are also important numbers required for a study. In this study, only 6.9% of the numbers telephone were disconnected. This grand total also showed some variation by sampling area, but deviations were overall not very great. Disconnected numbers, then, may be problematic, but were much less so than "no answers".

Those respondents who were still in school or who possessed at least a high school diploma were classified as "ineligible and were not interviewed. Results showed that approximately one-third (31.6%) of the telephone contacts fell in this category. Major differences were observed between SMSA dwellers (34.5%) and non-SMSA dwellers (22.4%), reinforcing the notion that rural areas may manifest lower levels of educational attainment than more urban areas. Even more variation, however, was observed across SMSA sampling areas. Percentages for example, ranged from a low of 18.1% to a high of 41.9%. Attempts to reconcile these differences by controlling for the actual number of "ineligibles" in an SMSA failed to resolve the differences. It remains unclear as to why this result should occur. At any rate, specialized populations like high school dropouts may be more or less difficult to locate by telephone depending on the presence of these "ineligibles".

Initial refusals by respondents which necessitate the need for oversampling may somehow bias the resulting sample. If initial refusals, for example, were all dropouts, then the sample of completed interviews would be greatly biased by virtue of having not included them. In this study initial refusals were substantial (10.7%). This was more the case for SMSA's (12.7%) than for non-SMSA's (4.4%). Again, substantial differences were noted across SMSA areas with a high of 18.4% to a low of 4.1%. This more than anything else probably represents variations in effective interviewing technique.

A final point of interest in Table 2 was the rate of completion calculated by dividing the number of telephone numbers used by the number of completed interviews obtained. The total completion rate of 13 to 1 shows that our initial estimate of 15 to 1 was extremely accurate overall, but that within some SMSA's and non-SMSA's this rate fluxuated significantly. As a result, sampling of this kind requires that analysts be prepared to draw additional samples from some areas in order to supplement low estimates or that generous estimates for all areas at the outset of the sampling process be undertaken.

Since little sampling bias was observed across sampling areas with regard to demographics, it was unlikely that the above factors themselves led to additional bias in the sample. However, because of their fluctuation across areas, this remains a potential problem in producing a representative sample. Most importantly for this study, the results illustrate the potential unpredictability of some factors in attempting to design a statewide sample.

Telephone Directory Publication Dates as a Factor

The date of publication of a telephone directory should greatly affect completion rates and sample representativeness because of invalid numbers and not yet published new numbers. This becomes even more problematic in a statewide survey where directories are likely to be published at different times throughout the year. Analysis attempted to assess the impact of date of publication on the sample.

Surprisingly, results showed that there were no significant differences with regard to completion rates, "no answers" and disconnected numbers. Furthermore, no significant variation was noted

for representativeness in the sample and date of directory publication. In this case, then, the expectation that the date of publication would be important was not borne out in the data.

Conclusions

Results of this analysis suggested two very general conclusions. First, the identification of special populations like high school dropouts may lead to sacrifices in terms of the representativeness of the completed sample. This occurs because so much "filtering" is required at the outset of the interview to locate a high school dropout that concerns for such things as selecting appropriate numbers by sex, race and age will lead to premature termination of the interview by potential respondents. As a result, one may have to settle for weighting data once interviews are completed.

Second, a reasonable amount of accurate prediction may be possible for a statewide sample design involving high school dropouts. But when one looks more closely at predictions about portions of the statewide sample design, major deviations from these initial predictions should be expected. So for example, similar SMSA areas may yield very different completion rates all things considered equal. Predictions about these differences seem to be difficult or impossible to account for prior to completing a study. Researchers must, therefore, develop contingency plans for dealing with these differences.

Footnotes

[1] Tabular data available from the author.

References

Boggs, David L., et al., Ohio Citizens Eligible for Adult Basic Education (Columbus: Ohio Department of Education, 1978).

Boggs, David L., et al., A Followup Study of Adult Basic Education in Ohio (Columbus: Ohio Department of Education, 1979).

Brunner, James A. and G. Allen Brunner, "Are Voluntarily Unlisted Telephone Subscribers Really Different?", Journal of Marketing Research 9 (1971) 121-124.

Dillman, Don, et al., "Reducing Refusal Rates for Telephone Interviews", Public Opinion Quarterly 40 (Spring 1976) 66-78.

Glasser, Gerald J. and Gale D. Metzger, "National Estimates of Nonlisted Telephone Households and Their Characteristics", Journal of Marketing Research, 12 (1975) 359-361.

Glasser, Gerald J. and Gale D. Metzger, "Random Digit Dialing as a Method of Telephone Sampling", Journal of Marketing Research, 9 (1972) 59-64.

Klecka, William R. and Alfred J. Tuchfarber, "The Efficiency, Biases and Problems of Random Digit Dialing", a paper presented to the 1974 Annual Conference of the American Association for Public Opinion Research.

Perry, Joseph B., "A Note on the Use of Telephone Directories as a Sample Source", Public Opinion Quarterly 32 (1968-1969) 691-695.

Sudman, Seymour, Applied Sampling (New York: Academic Press, 1976).

Sudman, Seymour, "On Sampling of Very Rare Human Populations", Journal of the American Statistical Association 67 (1972) 335-339.

U. S. Bureau of the Census, General Social and Economic Characteristics, PC(1).C37 Ohio (Washington, D.C.: GPO, 1972).

Table 1

Sampling Areas for a Statewide Survey of Ohio

Sampling Areas	Date Directory Published	Percentage of Population of Ohio	Required Completed Interviews
Area 1			
Cincinnati SMSA	June 1977	10.5%	157
Hamilton SMSA	October 1976	2.3%	34
Area 2			
Dayton SMSA	February 1977	8.2%	122
Springfield SMSA	June 1977	1.6%	23
Area 3			
Lima SMSA	September 1976	1.8%	26
Area 4			
Toledo SMSA	October 1976	5.6%	83
Area 5			
Columbus SMSA	July 1977	9.8%	146
Area 6			
Mansfield SMSA	November 1976	1.4%	20
Area 7			
Steubenville SMSA	October 1976	1.2%	17
Area 8			
Akron SMSA	December 1976	6.6%	98
Canton SMSA	June 1977	3.7%	55
Cleveland SMSA	May 1977	19.8%	296
Lorain SMSA	1977-1978	2.5%	37
Area 9			
Youngstown SMSA	March 1977	5.3%	79
Total SMSA		80.3%	1193
Total Non-SMSA		20.7%	310
Grand Total		100.0%	1503

Table 2

	Total Numbers Used (N)	Completion Rate	Initial Refusal (%)	Within Inter. Withdraw (%)	Non-Respond. Refusal (%)	Discon-nected (%)	Ineligible (%)	No Answers (%)
Cincinnati	2020	13/1	10.5	1.0	0.1	9.0	26.7	44.9
Hamilton	318	9/1	4.1	0.0	0.0	5.7	28.3	51.3
Dayton	1263	10/1	13.9	0.5	1.0	7.0	41.2	26.8
Springfield	304	13/1	7.6	0.3	0.3	7.2	31.9	45.1
Lima	250	10/1	18.4	0.4	0.0	2.0	28.4	40.4
Toledo	1000	12/1	11.6	0.4	0.3	7.9	32.1	39.4
Columbus	2451	17/1	10.3	0.7	0.1	9.5	34.3	39.1
Mansfield	265	13/1	5.7	0.0	0.0	9.1	18.1	59.6
Steubenville	325	19/1	17.8	0.0	0.0	6.5	31.1	39.4
Akron	1466	15/1	15.8	0.8	0.0	4.6	32.7	39.5
Canton	703	13/1	9.2	0.1	0.0	3.6	20.9	58.3
Cleveland	3412	12/1	15.4	1.3	0.3	7.8	41.9	27.5
Lorain	478	13/1	8.4	0.8	0.2	8.6	33.7	40.6
Youngstown	975	12/1	16.2	0.5	0.0	2.9	41.7	30.6
Total SMSA	15233	13/1	12.7	0.8	0.2	7.2	34.5	38.9
Total Non-SMSA	4678	15/1	4.4	0.2	0.0	6.0	22.4	59.3
Grand Total	19911	13/1	10.7	0.7	0.2	6.9	31.6	43.7