AN EXPLORATION OF COVERAGE IN FOUR DEMOGRAPHIC SURVEYS

Brian A. Harris-Kojetin, Arbitron, and Mick P. Couper, University of Michigan
Brian A. Harris-Kojetin, Arbitron, 9705 Patuxent Woods Dr., Columbia, MD 21046

Keywords: Within-household coverage; whole-household coverage; survey-census match

1. Introduction

Coverage error is usually defined as the discrepancy between statistics calculated on the frame population and the same statistics calculated on the target population (Groves, 1989, p. 83). Coverage error arises from the failure to give some units (households or persons) in the target population any chance of being included in the survey, or from the erroneous inclusion of ineligible units or duplication of eligible units. For many U.S. government surveys, the target population is defined as the civilian noninstitutionalized household population of the United States. In such surveys, information is sought about all members of the household, rather than just a single selected person. These surveys may fail to obtain information for certain household members because of their erroneous exclusion from the list of eligible household members. In other words, while there may be no explicit selection method, a portion of the eligible population is still missed through listing errors, rather than through nonresponse.

Coverage error is a function of both the proportion of the target population that is not covered and the differences between the covered and not covered populations with respect to the statistic of interest. Furthermore, the effect of noncoverage error on survey estimates may be compounded by other sources of error. For example, if the types of persons or households likely to be missed from the frame are also more likely to be nonrespondents, the combined effect on estimates may be large. It is often difficult to separate out the various sources of error, and comparisons of key estimates to external data sources may combine multiple error sources, potentially inflating the apparent effect of noncoverage. For example, the exclusion of certain household members may arise through interviewer error, through respondent misunderstanding of the question, through deliberate misrepresentations of household composition, and so on. Regardless of the cause, the net effect is still the same — the frame population is not the same as the target population.

There are two major sources of undercoverage in household surveys: that arising from the exclusion of whole households and that arising from the exclusion of eligible persons within sampled households.

Whole-Household Noncoverage. A number of stages in the survey process can produce coverage error. For example, the first step is usually the construction of the sample frame, which involves the listing and selection of housing units within a prescribed geographical area. For many surveys this is done once every ten years, using the decennial census as the basis for the construction of the survey frame with periodic updating in the intercensal period. To the extent that some housing units are erroneously missed during this step, or that new housing units have been added since the listing stage, these units are not included on the sample frame. New construction, mobile homes, other temporary quarters, and subdivided units may be disproportionately missed as the frame deteriorates over time.

Once the sampled housing units have been identified, the interviewer then visits the unit to determine if it is still occupied, and if so, seeks to identify and interview the occupants. During this stage, housing units may be misclassified as ineligible for a number of reasons. Some of these may be errors on the part of the interviewer (whether deliberate or accidental). Housing units in high crime areas inner-city neighborhoods, for example, may be disproportionately classified as vacant. Similarly, legally or illegally converted housing units may be disproportionately missed.

Within-Household Coverage. Once a unit has been identified as an occupied residence, that unit is deemed eligible for the survey, and the interviewer attempts to identify all eligible persons within the household. At this point the omission of eligible persons within the housing unit may occur. The source of such errors may be either the reporting person (the sample person providing the household roster information) or the interviewer, or both. Again, there may be a variety of reasons why certain persons are not included on the list of eligible household members, some deliberate and some accidental.

1.1 Estimates of Coverage Error

There are a number of ways to estimate the extent of coverage error. Much of the research on survey undercoverage has involved the comparison of aggregate survey counts of demographic subgroups to decennial census counts or the examination of poststratification factors. However, it is known that the decennial census is also subject to coverage error, and such comparisons may underestimate the true extent of the undercoverage. This is especially true if the same subgroups that are likely to be missed in surveys are those likely to be missed in the census. Furthermore, good estimates of survey coverage error are rare and costly (as noted for the

Canadian context by Clark, Kennedy and Wysocki, 1993). Much of the research effort in the U.S. has focused on census undercoverage rather than survey undercoverage; however, our focus here is on survey coverage. For the reasons mentioned above, survey undercoverage has not received as much research attention as many other sources of error in surveys (e.g., nonresponse error, measurement error).

Shapiro, Diffendal and Cantor (1993) compared aggregate data from the Current Population Survey (CPS) to 1980 decennial census control total data (adjusted using PES estimates) to explore issues of coverage in the CPS. They analyzed the coverage of household heads and other household members separately to produce estimates of whole-household and within-household undercoverage, respectively. They concluded that about 40% of the undercoverage in the CPS was due to whole-household misses, and 60% to within-household misses. The ratio of whole-household coverage to within-household coverage did not vary much for White and Black households. For the total population, they report an undercoverage rate of about 7.7%. However, there are large differences in the overall rates of undercoverage by race: for Whites, the overall rate was 6.8% while for Blacks it was 17.1%. In other words, for Blacks both the whole-household and within-household coverage was larger than for Whites.

Fay (1989) reports on a similar analysis, but focusing on within-household coverage in the 1980 CPS, using a direct match of CPS sample persons to the 1980 decennial census in CPS interviewed households. He presents separate analyses for persons 25 and older, persons 15-24 and persons under 15. The primary reason for this is that the different definitions of household membership (particularly for students in college dormitories) between the survey and the census make direct comparisons across all age groups difficult. Black males age 25-44 had the highest undercoverage rates of this age group in 1980. For persons 15-24, it appears that Hispanic males have the highest undercoverage rates. Fay (1989) also examined the net within-household undercoverage by various household characteristics. He found these rates to be highest for non-relative males (22.7%) and other male relatives (17.8%), higher in central cities (2.7%) than non-SMSAs (1.6%), and higher for renters (2.7%) than owners (1.6%).

1.2 Research on Noncoverage

We have seen that there appears to be differential undercoverage by certain key demographic variables (age and race, among others). A variety of different approaches have been used, both to understand the causes of survey undercoverage and to reduce undercoverage. Some of these approaches focus on whole-household undercoverage, while others focus on within-household undercoverage, and still others on both. The two major approaches involve ethnographic research and rostering research aimed at uncovering missing persons within households.

The matching of survey cases to decennial census data from the same households is one way to explore within-household undercoverage that overcomes some of the limitations of these other approaches. However, such match studies are rarely undertaken. One such study was conducted on the CPS, with matching to the 1980 decennial census (see Fay, 1989). In this paper we report on a more recent effort, based on the 1990 census, designed primarily to explore survey nonresponse (see Groves and Couper, 1993, 1998). While our data were not designed for this purpose, we nonetheless conduct an exploratory analysis of coverage issues to demonstrate the potential utility of using matched data in this way. With the next decennial census approaching, we feel that now is a particularly important time to raise these issues. Almost two decades have passed since the last such study focused on noncoverage, and this examined only a single survey.

2. Methods

The purpose of the present study was to utilize the data from the 1990 survey-census match study to gain insight into correlates of noncoverage across four major demographic surveys conducted by the Census Bureau. Despite the limitations of these data, they offer a rare opportunity to explore the potential for coverage research using similar approaches in surveys.

2.1 Data Sources and Census Matching

The coverage analyses are based on survey data matched to the 1990 decennial census. This section briefly described the data used and the process of the match operation. The survey-census match was conducted to investigate survey nonresponse (see Groves and Couper ,1998), and was not originally designed for coverage analysis. While seven surveys were used in the match study, the coverage analyses focus on only four ongoing major demographic surveys conducted by the Census Bureau: Consumer Expenditure Survey (CEQ), Current Population Survey (CPS), National Health Interview Survey (NHIS), and National Crime Survey (NCS). For the nonresponse study, all nonrespondent cases and a sample of respondent cases from the 3-month period surrounding census day (April 1, 1990) were matched to decennial census records. We used only the interviewed cases here.

The number of interviewed cases included in the match study for each of the four surveys is shown in Table 1. It is important to note that the match was an

address (rather than household or person) match. For most of the nonresponse cases, we had no information other than the sample address. Where additional information was available (such as household composition or name, primarily in the case of interviewed cases), this information was used to verify the quality of the match. The match itself was a clerical operation, using survey segment listing and sample unit identifiers (addresses, descriptions, etc.) to look up the corresponding address in the decennial census Address Control File (ACF). Several additional steps were undertaken for cases that did not produce an exact match (perfect correspondence between the survey and census address, including all prefixes and suffixes and unit designations).

The percentage of interviewed cases successfully matched at the address level for the four surveys is also shown in Table 1. Over the four surveys included in the present study, 98.4% of interviewed cases were matched to the census address. (These match rates are after excluding a small number of group quarters cases from each survey.) Of the remaining cases, some were matched to several units at an address (i.e, the building was identified, but the exact unit could not be determined), some were matched to the census block, but the unit could not be identified, and a small number were not successfully matched at all or were not attempted.

The coverage analyses are based on the 98.4% of cases that were successfully matched to the census address for the four surveys. Even a match at the address level, however, does not guarantee that the same persons occupied the housing unit in the survey and census enumeration. Selecting survey cases interviewed around census day reduces the likelihood of households changing. However, in the few cases where this may have occurred, it is reasonable to assume that, given the high level of geographical clustering of certain characteristics (such as race and household size) in the U.S., the households that moved in are likely to resemble the households that moved out in terms of such characteristics.

A quality control procedure was implemented to review a sample of 10% of all cases matched. Of the 2,057 cases selected for review, 1,361 contained last name and/or household composition information on the survey form. Of these cases where some information was available, 91.1% were found to match on last name or household composition or both. In the remaining 8.9% it was unclear whether the same household was captured, or insufficient evidence was available to make a judgement about the match of household members.

There could be several reasons for the non-matches. One could be that the address was recorded incorrectly on either the survey or the census form. This appeared to be a particular problem in multi-unit structures where the unit designators (apartment numbers) are unclear. Other errors could occur on either the census (e.g., missed unit, incorrect geocoding, dwelling incorrectly classified as vacant, etc.) or survey address listing (transcription errors, wrong housing unit interviewed, etc.). However, we are confident that the high success of the match operation and the rigorous quality control procedures instituted, ensured a high match rate, and a high degree of correspondence between the census and survey information for matched units.

Despite the limitations of these data, and the fact that they were not collected with coverage analyses in mind, we believe these data are nonetheless useful for exploratory analysis of issues related to survey coverage.

2.2 Measures

A variety of independent variables were drawn from the census data. All of these were at the household level or higher, and include indicators of urbanicity, household size, the number of housing units in the structure and tenure of the occupants (owner or renter). Some person-level characteristics were aggregated to the household level, or household-level equivalents of person characteristics were used, e.g., race of householder or reference person, household composition, presence of persons of certain ages in the household, and so on. The household (or more correctly the address) is the unit of analysis.

2.3 Overview of Analyses

All analyses reported here were based on the combined dataset across all four surveys. Some initial comparisons were conducted on the surveys individually and the pattern of findings was very similar across surveys, so only these aggregate results will be presented. All of the analyses reported are based on weighted data. The weights are a product of the original selection weights for the survey and the selection weights for inclusion in the survey-census match project. The relative weights of the four surveys are such that each survey should contribute about equally to the combined estimates, given minor differences in definitions of population. The standard errors and statistical tests reflect clustering in addition to the weights and were conducted using SUDAAN.

2.4 Limitations

As noted already, the match data were not designed with coverage analyses in mind, and suffer from several limitations. Because the focus was on nonresponse, nonrespondent cases were oversampled. Furthermore, the match was conducted at the household or address level rather than the person level. This means that no attempt

was made during the match to ensure that the people listed on the census form were the same as those on the survey roster. In most cases, names were not available to the match clerks at the time. Thus, in some small percentage of cases, different families will have been captured by the census and the survey, but will still be treated as having matched.

In addition, we cannot assume that the decennial census counts are the gold standard. Strictly speaking, we should thus talk about differential coverage, or discrepancies between two methods of enumerating household members. Both sources of household counts are subject to error, and these may be errors of different kinds (the census being largely self-enumeration, and the surveys being interviewer-administered). Furthermore, we are comparing survey counts to raw census counts, which we know from the Post-Enumeration Survey (PES) underestimate certain groups such as minorities.

Despite these and other limitations, we believe these data are suitable for exploratory analysis of differential coverage. Our main goal is to explore the feasibility and utility of these kinds of data and analyses for the investigation of survey coverage.

3. Results

The analyses were conducted in three phases. First, comparisons were made at the household level for survey over- and undercount relative to the census. Next, bivariate analyses were conducted examining the characteristics from the decennial census that were related to survey over- and undercoverage. Finally, because there was overlap among the characteristics examined, we entered all the census characteristics into multinomial logistic regression models to identify the unique predictors of survey under- and over-coverage.

3.1 Estimates of Over- and Under-coverage

We initially examined the full distribution of relative coverage of household members between the two sources. Given that the distribution of the difference in household counts peaks at 0 (no difference) and falls off steeply at higher numbers, we created a collapsed measure that simply indicates whether the survey counts are higher, lower or the same as the census counts. Table 2 shows the *unweighted counts* of the number of households that were matched to the 1990 decennial census, and whether the survey had fewer household members (undercoverage), more household members (overcoverage) or the same as the Census. It is again important to note that the census counts do not necessarily represent "truth." It is apparent that there is slightly more relative survey undercoverage than overcoverage, but the distribution is fairly symmetric.

3.2 Bivariate Analyses of Relative Household Coverage

As can be seen in Table 3, there are a variety of characteristics related to both survey undercoverage and overcoverage. There is greater relative undercoverage of persons in housing units occupied by renters, in small multiunit structures, and households in central cities compared to single family homes, those occupied by owners, and those in rural areas. Survey undercoverage also appears to increase with increasing household size. There is greater undercoverage for households headed by Blacks or Hispanics compared to households headed by whites and other racial groups, and less undercoverage of households consisting of just nuclear families or persons all over the age of 70.

There is greater relative survey overcoverage of persons in housing units occupied by renters, in multiunit (2-9 unit) structures, located in central cities compared to single family homes occupied by owners in rural areas. Overcoverage appears to be highest for single person households and decreased with increasing household size. There is greater overcoverage for households headed by Blacks or Hispanics compared to households headed by whites and other racial groups, and less overcoverage of households consisting of persons all over the age of 70.

3.3 Multivariate Analyses of Relative Coverage

Because there is overlap among several of the variables, multivariate analyses were also conducted to identify more clearly the unique contributions of these variables to under- or overcoverage. Table 4 contains the results of multinomial logistic regression predicting the three-way coverage variable. This model is the result of several model-fitting activities, in which other variables were included and alternative variable formulations were tested. The estimated odds ratios presented in Table 4 reflect both the weights and the clustered design.

As can be seen in Table 4, several household composition variables (household size, nuclear households, presence of young children) remain statistically significant predictors of survey undercoverage in the multivariate model. Specifically, nuclear households were about one fourth as likely to exhibit survey undercoverage, that is, to have fewer household members counted in the survey than in the census. The presence of young children similarly reduces the odds of survey undercoverage. We are surprised by the effects of race/ethnicity on coverage in the multivariate model. Several additional analyses have not produced plausible explanations for these results.

As can also be seen in Table 4, a number of census variables were related to survey overcoverage, that is, higher counts of household members in the survey than in the census. Specifically, the race of the householder, the number of units in the structure, the age of the occupants

and whether the occupants own or rent the housing unit are related to survey overcoverage. More specifically, relative to households occupied by owners, renters are nearly twice as likely to have overcoverage. Overcoverage is also more likely to occur in all housing structures containing less than 10 units than those containing more than 10 units. Households headed by Hispanics are nearly 4 times as likely to have some level of survey overcoverage. Finally, relative to households with all persons greater than 70 years of age, all other households are more likely to have survey overcoverage.

4. Conclusions

We offer the present study as an illustration for further research on coverage. Because the current dataset was not explicitly designed for the study of coverage issues, it suffers from a number of limitations. Nonetheless, a survey-census match study such as this offer great promise as a vehicle for understanding the extent of coverage problems and possible correlates and causes of coverage errors. We believe greater research attention needs to be paid to coverage errors, and in that vein offer suggestions for a research agenda below.

Taking the above idea one step further would involve the collection of micro-level data at the household or person level. This would involve the matching of rostered or interviewed survey households to decennial census data at the time of a decennial census. We believe that survey-census match studies offer promise for exploring issues of noncoverage, both whole-household and within-household. The last such study designed to explore noncoverage was conducted on the CPS in 1980 (Fay, 1989). With the next decennial census fast approaching, we urge survey sponsors to consider repeating such a study on a large scale. Such a study could focus on housing units classified as ineligible in the survey to examine issues of whole-household undercoverage. In addition, interviewed cases could be matched to decennial census data to explore issues of within-household noncoverage. Efficiency could be maximized by oversampling cases in areas with higher than average expected rates of survey undercoverage. Depending on the final enumeration strategy chosen for the census, it may be useful to restrict the match to areas sampled for intensive follow-up in the census, if these overlap with survey PSUs. This would ensure that the census count of household members is the most accurate possible.

In conclusion, while the data at our disposal were not designed to explore survey coverage issues, we have demonstrated that such an approach yields useful information for the study of survey coverage. The results we obtained generally match those from other studies of survey coverage conducted in the last several decades. Despite their limitations, match data offer several advantages over other approaches to studying coverage; they permit comparative analysis across several surveys; they permit exploration of a relatively large number of correlates of coverage; and facilitate the examination of several correlates simultaneously through multivariate analysis. Survey undercoverage has generally suffered from a lack of research attention. The upcoming decennial census in 200 provides an ideal opportunity to explore issues related to survey coverage.

References

Clark, C., Kennedy, B., and Wysocki, M. (1993), "Coverage Error in the Canadian Labour Force Survey." *Proceedings of the Bureau of the Census Annual Research Conference*, pp. 620-637.

Fay, R.E. (1989), "An Analysis of Within-household Undercoverage in the Current Population Survey." *Proceedings of the Bureau of the Census Annual Research Conference*, pp. 156-175.

Groves, R.M. (1989), *Survey Errors and Survey Costs*. New York: Wiley.

Groves, R.M. and Couper, M.P. (1993), "Unit Nonresponse in Demographic Surveys." *Proceedings of the Bureau of the Census Annual Research Conference*, pp. 593-619.

Groves, R.M. and Couper, M.P. (1998), *Nonresponse in Household Interview Surveys*. New York: Wiley.

Shapiro, G.M., Diffendal, G., and Cantor, D. (1993), "Survey Undercoverage: Major Causes, and New Estimates of Magnitude." *Proceedings of the Bureau of the Census Annual Research Conference*, pp. 638-663.

Table 1. Unweighted Match Results

Survey	Matched Households	Match Rate
CEQ	830	99.9%
CPS	1,397	99.6%
NHIS	1,800	96.7%
NCS	1,316	96.5%
Total	5,343	98.4%

Table 2. Summary Counts of Differential Coverage by Survey (Unweighted percentages)

Survey	Survey Under-coverage	Same as Census	Survey Overcoverage
CEQ	8.0	84.9	7.1
CPS	8.7	82.5	8.7
NHIS	12.6	82.1	5.4
NCS	10.3	84.7	5.0
Total	10.3	83.3	6.4

Table 3. Relative Coverage for All Persons in Household (in percent)

	Survey Under	Same	Survey Over
Overall	8.9	85.5	5.6
Race/Ethnicity			
Hispanic	16.2	69.5	14.3
Black	13.5	79.7	6.9
Other	7.7	87.4	4.9
$X^2 = 22.6$, df=4, p<0.001			
Household Composition			
Nuclear household	3.6	91.0	5.5
All other	22.1	72.1	5.8
$X^2 = 102.4$, df=2, p<0.001			
Age Composition			
All under 30	9.8	84.4	5.7
Mixed	9.7	84.2	6.1
All over 70	1.7	96.3	2.0
$X^2 = 59.6$, df=4, p<0.001			
Sex Diversity			
Single adult in HH	2.6	90.3	7.1
All adults of same sex	21.6	70.5	8.0
Adults of different sex	10.5	84.8	4.7
$X^2 = 70.8$, df=4, p<0.001			
Children Under 5 in HH			
Yes	11.8	84.4	3.8
No	8.3	85.7	6.0
$X^2 = 8.2$, df=2, p=0.017			
Units in Structure			
Mobile home	9.3	83.3	7.4
Single family	7.8	87.4	4.8
2-9 units	13.2	76.9	9.9
10 or more units	9.4	85.8	4.8
$X^2 = 16.0$, df=6, p=0.014			
Household Size			
One	0.1	91.8	8.1
Two	6.1	88.8	5.1
Three	13.5	81.6	4.9
Four	11.3	84.0	4.6
Five or more	22.4	72.7	4.9
$X^2 = 178.2$, df=8, p<0.001			
Tenure			
Own	7.6	88.0	4.4
Rent, Other	11.4	80.6	8.0
$X^2 = 16.9$, df=2, p<0.001			
Urbanicity			
Central city	13.0	79.8	7.2
Suburbs	10.4	82.9	6.6
Other urban	8.7	86.6	4.7
Rural	6.2	88.5	5.4
$X^2 = 19.3$, df=6, p=0.004			

Table 4. Odds Ratios from Multinomial Logistic Regression Model

	Survey Under-coverage	Survey Over-coverage
Household Size	1.73***	0.82*
Race/Ethnicity		
Hispanic	1.06	3.75***
Black	1.17	1.42
Other	—	—
Units in Structure		
Mobile home	1.06	3.40**
Single family	0.72	2.08**
2-9 units	1.26	2.48***
10 or more units	—	—
Urbanicity		
Central city	1.34	1.09
Suburbs	1.47	1.16
Other urban	1.34	0.78
Rural	—	—
Household Composition		
Nuclear household	0.23***	0.77
All other	—	—
Age Composition		
All under 30	1.73	3.20**
Mixed	1.96*	4.38***
All over 70	—	—
Children Under 5 in Household		
Yes	0.58**	0.69
No	—	—
Tenure		
Own	0.89	0.50**
All other	—	—

* p<.10, ** p<.05, *** p<.01