CHARACTERISTICS OF CENSUS ERRORS

Deborah H. Griffin and Christopher L. Moriarity
Bureau of the Census, Washington, DC 20233

1. INTRODUCTION

This research project investigates the characteristics of coverage errors in the 1990 Decennial Census. Data which were collected in the Post Enumeration Survey (PES) and the Housing Unit Coverage Study (HUCS) provide the framework for the analysis. Data from the HUCS are not summarized in this report but will be analyzed over the next several months.

The results presented in this paper are based on PES results and therefore focus on those factors that might cause persons to be enumerated in error in the census. Our analysis is limited to persons living in housing units. There are two general categories of coverage errors - enumeration errors (persons we enumerated in error) and omissions (persons we erroneously did not enumerate). It is very important for us to understand where errors occur. Only then can we identify the causes of census errors and determine where we need to make changes. This investigation may suggest the need to revise census procedures or redesign census questionnaires.

This paper analyzes characteristics of enumeration errors. Enumeration errors include persons who were duplicated, persons who were counted in the wrong census geography or at the wrong address, fictitious persons, and other persons who should not have been included in the census. We also plan to analyze census omissions and the factors that might cause persons to be missed in the census. We intend to issue those results in a future report.

Census coverage error, which is nonsampling error, can be introduced at various stages of data collection and processing. Response error, interviewer error, and procedural and design errors all contribute to the total error recognized by the PES. This report examines the enumeration errors identified by the PES to determine if rates varied by:

- how the data were collected,
- who provided the data,
- when the data were collected, or
- the type of household or address.

2. METHODOLOGY

The data used in this report come from several sources - PES files, census files, and census questionnaires.

2.1 PES Data

PES data were used to identify census errors. The PES is designed to measure net error. The PES methodology identifies both over and under enumerations and combines these data to produce dual system estimates of net coverage error. To accomplish this, the design required the identification of "search areas" and the definition of "sufficient data for matching". Some of these techniques, although critical to the PES, complicate our analyses. Our research independently studies these two groups of errors. The PES was not designed for that type of analysis and therefore using the PES leads to some limitations in this research.

PES files include a set of codes that represent the "PES status" of persons who were enumerated in the 1990 Census. The "PES status" is essentially the conclusion from the PES of whether the person was correctly or erroneously enumerated in the census. The set of codes are detailed and allow us to categorize persons as:

- correctly enumerated,
- enumerated at the correct address that was coded to the incorrect geography,
- fictitious,
- duplicates, and
- other erroneous enumerations.

In some cases, a duplicate code was arbitrarily assigned to one of the two matching census persons. We made revisions to a small proportion of the codes so that our estimates of duplicates included a contribution from both of the matching census persons.

The erroneous enumeration data used in these analyses are those that were current as of December, 1991. The data differ slightly from the PES "production data", those that were used in mid-1991 to produce published estimates of the coverage of the 1990 Census. A small number of sample cases that were incorrectly assigned a final PES status of correctly enumerated due to a processing error during PES production were reassigned a revised final status of erroneously enumerated. Our analysis uses the revised final status.

In addition, a small percentage of cases (approximately 1.2 percent) did not have sufficient information to allow for a determination of enumeration status. We refer to these cases below as insufficient information (II) cases.

In this report we refer to all differences between the PES and the census as "census error". We know, from evaluation studies we have conducted [1] [2], that some proportion of these reflect measurement error in the PES rather than census enumeration error. The PES errors occur due to response error, matching error, and processing error just like in the census. Even though we cannot disentangle the two sets of errors, we think it is

instructive to look at the combined set to see if there are trends or other anomalies that help us understand the census and the PES data collection and processing.

2.2 Census Data

Census data were used to summarize the characteristics of persons, addresses and households that were enumerated in error. During the census, automated data were maintained that provide a substantial amount of information on who provided the census data and when and how the data were collected. Critical data that were not available from either the census or PES files were obtained for this study from a clerical review of the actual census questionnaires from the PES sample areas. Information such as who completed the questionnaire and when it was completed was clerically coded and keyed.

2.3 Estimation

We linked data from the three sources (PES data, census data, clerical data) and produced weighted estimates of erroneous enumeration rates for specific characteristics. We omitted the II cases prior to computing any estimates of erroneous enumeration rates. This is equivalent to assuming that the erroneous enumeration rates for II cases are the same as the erroneous enumeration rates for the other cases.

We computed design-based stratified jackknife estimates of standard errors using VPLX, a general-purpose variance estimation software package developed by Robert E. Fay, Senior Mathematical Statistician at the Census Bureau. We used these standard errors to produce confidence intervals and conduct hypothesis tests. We performed all hypothesis tests at a significance level of 10 percent. We did not employ a multiple comparison methodology for our hypothesis tests.

3. RESULTS

3.1 Overall

We estimate with 90 percent confidence that between 4.3 percent and 4.8 percent of the persons enumerated in the census were enumerated in error. Approximately 1.7 percent of all enumerated persons were found to be duplicated. Most of those were "within block" duplicates (1.2 percent). The remaining 0.5 percent were duplicated outside the block but within the PES Search Area. Other errors represented the largest component (2.4 percent). This category includes all persons determined by the PES to have been enumerated at the wrong address. Persons were identified as fictitious at a rate of 0.2 percent. The final category, geocoding errors, summarizes only those persons enumerated at addresses that were geocoded in error outside of the PES Search Area (0.4 percent).

3.2 Mail Return Households

It is of particular interest to investigate the enumeration errors for persons who were enumerated on questionnaires that were completed and returned in the mail. Although most mail return questionnaires are completed by household members, on occasion we will receive and process mail return questionnaires that are completed by proxies such as a relative or landlord. Therefore errors on mail returns are largely response errors or errors caused by our delivery or address list development procedures.

In this study all persons who were listed on questionnaires that were used to record mailback responses are classified as members of mail return households. Persons listed on "enumerator friendly questionnaires", which were used in the followup activities, are classified as members of enumerator completed households. Some persons on mail return questionnaires may have been added to these households as a result of an edit followup or search/match operation. This study does not attempt to identify those persons or assign separate error rates.

When error rates are high for mail return households it suggests that response error is a factor that must be addressed for the 2000 Census. We depend on a respondent correctly completing a questionnaire. We do however recognize that respondents may misunderstand who to include as part of their census household. Respondents may fail to read the instructions or may become confused with the concept of "usual residence". Respondents may also intentionally exclude persons from their census forms or include persons that they realize should not be included.

In our study approximately 73.9 percent of all persons were enumerated on mail return questionnaires. We estimate with 90 percent confidence that between 2.9 and 3.3 percent of these persons were enumerated in error.

Graph 3.2 displays the distribution of these errors. Approximately 54.1 percent of these errors were persons who were counted at one address when they should have been counted at another (for example, the PES confirmed that they moved in after census day or were not a household member according to census residence rules) or not counted at all (for example, they were born after or died before census day). Duplicates describe those situations when the PES confirmed that the person was also enumerated at another address. About 32.5 percent of the persons on mail returns who were enumerated in error were determined by the PES to be duplicates (21.3 percent - within block, 11.2 percent - outside of block). It is possible that some duplicate enumerations represent substitutions for the correct enumeration. In some instances we may enumerate one three- person household twice and miss another

three-person household. We do not know the extent to which this may have occurred. Approximately 1.7 percent of the errors were classified as fictitious persons. The final category, geocoding errors, describes persons who were enumerated at the correct address when that address was incorrectly geocoded (11.8 percent).

Other errors and duplications make up nearly 87 percent of all enumeration errors on mail returns. Duplications would occur if a household received and completed two or more questionnaires. Delivery errors could result in duplication if the questionnaire was delivered in error, the household completed the form and due to a mixup during nonresponse followup the household was enumerated again. "Other errors" are likely to be response errors due to misunderstandings about census residence rules and who should be included as a part of the census household. These errors may be due to a poorly designed questionnaire. Given that this type of error accounts for over one half of the persons who were enumerated in error on mail returned questionnaires, strong consideration should be given to research into clarifying the definitions of who should be included as part of the census household. It is also possible that the estimate of "other errors" is high due to recall bias during PES followup. If, during followup, a respondent incorrectly recalls where they lived on census day the PES might erroneously conclude that the household was enumerated in error in the census.

It is valuable to study how these rates varied by certain characteristics. Sections 3.2.1 - 3.2.6 provide this additional detail.

3.2.1 Characteristics of the Respondent

A review of the back page of the mail return questionnaires allowed us to identify who the respondent was for each household. A respondent's name was provided for about 88.1 percent of the mail return questionnaires in our sample. From this information we could determine if the respondent was a member of the household or a proxy. We also could profile other characteristics of the respondent.

Approximately 99.7 percent of the mail return questionnaires identifying the respondent were completed by a household member, the remaining 0.3 percent having been completed by a nonhousehold member who could have been either a relative, a neighbor or a landlord. The erroneous enumeration rate for persons enumerated on questionnaires completed by a household member was 3.0 percent while the rate for persons enumerated on questionnaires completed by a proxy was 7.0 percent. This suggests that the preferred respondent is a household member. Proxies are more likely to include persons who should not be enumerated at that address. It is not surprising that proxies would have less information to provide the most correct responses. This may also be due to differences in error rates for varying types of households, especially those with a large number of nonrelatives.

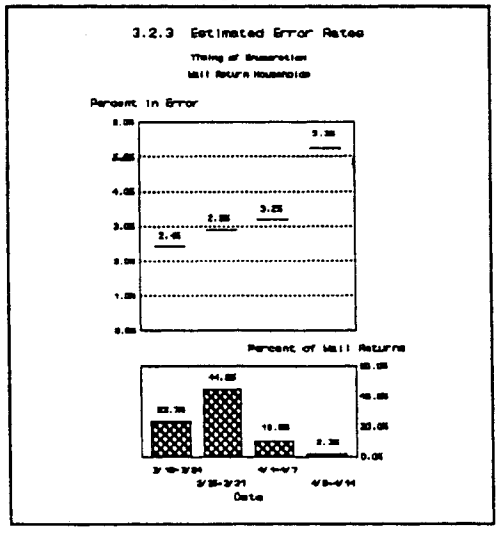
Analysis of data summarizing various demographic characteristics of the respondents indicates that there does not appear to be any effect on the rate of erroneous enumeration due to either the sex or the age of the respondent.

3.2.2 Form Length

Persons who were enumerated on long form mail return questionnaires had no significant difference in the rate of error from those enumerated on short form mail return questionnaires. They also appear to exhibit consistent distributions of types of errors. This suggests that respondents are equally likely to include persons in error regardless of the length of the form that they receive.

3.2.3 Timing of Enumeration

The dates of enumeration used in these analyses are based on annotations on the census questionnaires. On the back cover of the questionnaire respondents were asked to record the date of completion. Approximately 88.4 percent of all mail return questionnaires in our study provided an enumeration date. Analysis of the enumeration dates provided by respondents indicates that almost all mail returns were completed between March 20 and April 15. Graph 3.2.3 displays error rates for four successive weeks. Note that the rate of erroneous enumerations steadily rises from 2.4 percent to 5.3 percent in a four week time frame. These data suggest that the quality of data received from mail return households deteriorates over time. This graph also puts the impact of these errors in perspective by displaying the percent of the total mail return universe that were completed during these weeks. Note that only 2.3 percent of all mail returns had an error rate of 5.3 percent.



This decline could be due to several factors. It is possible that in-movers represent some of this universe. If a household moved in after April 1 and completed the questionnaire nonetheless, they would be erroneous enumerations. Finally the householders who are most conscientious about completing the form immediately might also be more conscientious in carefully reading the instructions and thus were less likely to include persons on the questionnaire that should not have been included.

When we look more closely at the distribution of types of errors over time we see that most types of errors occurred at a relatively constant rate (see Graph 3.2.3 below). Only "other errors" and "within block duplicates" rose over time.

3.2.4 Size of Household

We also analyzed error rates by household size. Rates were estimated for household sizes of 1, 2, 3, 4, 5, 6, 7, and 8 or larger. Persons who were enumerated to be the only person in the household were enumerated erroneously 2.9 percent of the time. Significant differences do not exist for most of these categories. The largest households, however, (size 5, 6, 7 and 8+) had a slightly higher rate of erroneous enumerations when compared to households of size 1, 2, 3 and 4 (2.9 percent versus 3.7 percent). In addition, mail return households with 8 or more persons had a higher erroneous enumeration rate (5.8 percent) than all other households (3.1 percent).

This may indicate that some form of edit may be appropriate for large households as mail respondents have a tendency to include persons in error at a higher rate.

3.2.5 Type of Structure

On the census questionnaire respondents are asked, "What best describes this building?". Response options include

- a mobile home or trailer
- a one-family home detached from any other house
- a one-family house attached to one or more houses
- a building with two apartments
- a building with three or four apartments, etc.

For these analyses we collapsed the first three categories into "single units" and the remaining categories into "multi-units". We plan to also look at the more detailed categories.

Approximately 17.3 percent of the persons in our sample who returned a questionnaire by mail were in multi-unit structures. The rate of error for persons in multi-units was 4.0 percent versus 2.8 percent for persons in single units. This could be due to a greater likelihood of delivery errors and apartment mixups in multi-unit structures. It may also be a function of the persons who live in multi-units and their likelihood of being enumerated in error.

3.2.6 Tenure

Analysis of tenure data for mail return questionnaires identifies renters as having a higher rate of erroneous enumerations. Approximately 2.8 percent of persons in owner occupied units (as opposed to 3.9 percent of persons in rental units) were enumerated in error. Again, this may be less a function of rental units and more a function of the persons living in rental units.

3.3 Enumerator Completed

During nonresponse followup, or another followup operation, an enumerator may misunderstand the procedures, including the concept of "usual residence". The enumerator may not reference April 1, and erroneously enumerate post-census day movers. An enumerator could become confused in a multi-unit structure and enumerate the wrong household. Enumerators could also intentionally fabricate (or "curbstone") data or bias responses by rewording questions. Response errors can also occur on enumerator filled forms when a household respondent supplies incorrect information to an enumerator. We estimate with 90 percent confidence that between 8.2 and 9.2 percent of the persons enumerated on questionnaires completed by an enumerator were erroneous enumerations.

The major causes of error for persons on enumerator completed questionnaires were other errors (49.9 percent) and duplication (40.2 percent). Within block duplicates had and error rate of 29.5 percent. Geocoding errors and fictitious persons were the cause of error 4.6 percent and 5.4 percent of the time.

Duplications can occur when enumerators visit the wrong address in nonresponse followup. Undoubtedly apartment mix-up and questionnaire delivery problems contribute to duplicates. "Other errors" could result if enumerators fail to reference April 1 during the nonresponse followup interview. "Other errors" will also occur when an enumerator does not understand who to include in the census day household.

Sections 3.3.1 - 3.3.5 summarize erroneous enumeration rates by certain characteristics of these enumerator completed households.

3.3.1 Type of Followup Procedures

During the final stages of nonresponse followup the district offices (DO) were instructed that they could implement "last resort and closeout" procedures to enumerate the final set of unresolved cases. "Last resort" procedures allowed the enumerators to turn in questionnaires with some items unanswered. "Closeout" procedures allowed questionnaires to be accepted with even less data. Critical information on household size and occupancy status was still required. Enumerators were

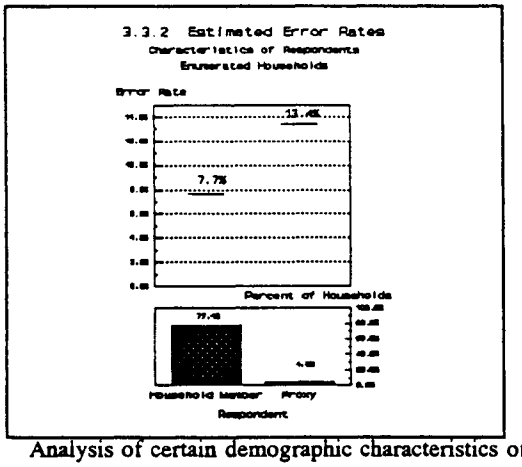
requested to code the questionnaires to record when they implemented these special procedures. The estimates below are based on these clerical codes and thus have potential limitations. The data show that persons who were enumerated using either last resort or closeout procedures have a significantly higher rate of erroneous enumerations than those enumerated under regular procedures (12.8 percent versus 8.4 percent).

Approximately 93.4 percent of all persons were enumerated on forms using regular enumeration procedures. About 6.6 percent were enumerated on forms coded as having been enumerated using either last resort or closeout procedures.

3.3.2 Characteristics of Respondents

We estimate that about 77.4 percent of the enumerator completed questionnaires were completed based on an interview with a household member. About 4.6 percent were completed based on information from a proxy such as a relative, landlord or neighbor. The remaining 18.0 percent cannot be classified. About 7.7 percent of the persons on questionnaires which were completed by an enumerator based on an interview with a household member were determined to be erroneous enumerations.

When the information came from a proxy this rate rose to 13.4 percent. These rates are significantly different. As with mail return households we confirm that household members are the preferred source for obtaining data. During followup activities, this is especially true.



Analysis of certain demographic characteristics of respondents indicates that neither the age nor the sex of the respondent impacted the erroneous enumeration rate. A review of the distribution of error types across various demographic groups did not detect any noteworthy differences.

3.3.3 Form Length

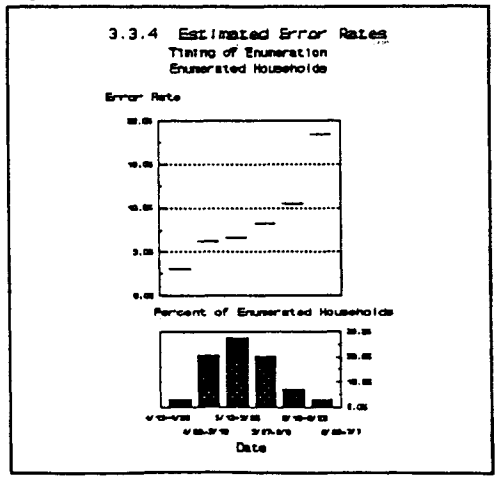
No difference was detected in the rates of error by form length for mail return questionnaires. This was not true for questionnaires that were completed by enumerators. Persons on short form questionnaires had an error rate of 9.3 percent while persons on long form questionnaires had an error rate of 6.5 percent. These rates are significantly different.

The higher error rate for short forms may be due to the fact that short forms may have been used in cases that were hard to enumerate. If an enumerator was having trouble making contact to complete an interview it is highly unlikely that he/she would have filled out a long form questionnaire. It is also likely that closeout and last resort interviews, having higher error rates, were recorded on short forms.

3.3.4 Timing of Enumeration

The date of enumeration is based on annotations on the census questionnaires. Enumerators entered the enumeration date on the front cover. These dates were coded for the analyses in this section. Nearly 92 percent of all enumerator return questionnaires provided an enumeration date.

The timing of nonresponse and field followup varied by DO. Most offices began nonresponse followup on May 3, but offices that were expected to have the greatest followup workloads began on April 26. Our data show that most enumerator returns were completed between April 26 and July 13. Graph 3.3.4 summarizes the rate of erroneous enumerations on enumerator completed questionnaires over time. Six two week intervals are shown. The rates increase from approximately 3.1 percent error to 18.4 percent. Many factors could contribute to this increase in erroneous enumerations over time. Household composition changes, there is an increase in the impact of movers, and the hardest persons to enumerate are likely to be enumerated in the final stages of the census. Recall error is also a likely contributor. Note that, as with mail returns, the time periods with the highest error rates had relatively low representation in the universe of enumerated households. Only 3.3 percent of all enumerator returns were completed between June 23 and July 7, the time period having an error rate of 18.4 percent.



3.3.5 Size of Household

We also analyzed the erroneous enumeration rates on enumerator completed forms by household size. As with mail return households, these rates were estimated for households of 1, 2, 3, 4, 5, 6, 7, and 8 or larger. These data suggest that the highest rates of error are found in the smallest households, although tests do not show all differences to be significant. One person households had a significantly higher erroneous enumeration rate (10.6 percent) than all other households (8.5 percent). This finding may confirm earlier research on "POP-1" cases which led to additional checks on one person households in nonresponse followup. Enumerators may have a harder time correctly locating and enumerating small households as they are harder to contact. Enumerators may also fabricate small households when they cannot make contact.

3.3.6 Type of Structure

Approximately 9.7 percent of the persons enumerated in multi-unit structures by an enumerator were enumerated in error. The rate of error for persons enumerated in single units was 7.5 percent. These rates are significantly different.

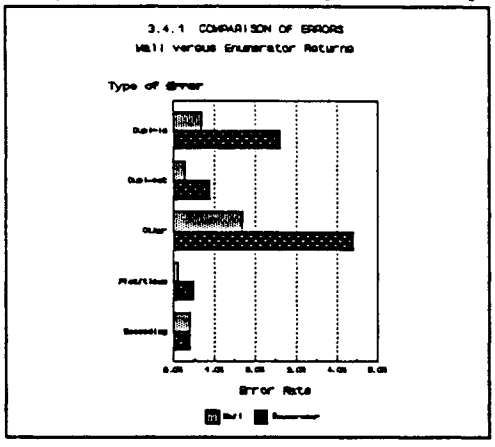
3.3.7 Tenure

Analysis of tenure data identifies renters who were enumerated during followup activities as having a significantly higher rate of erroneous enumerations than owners (9.6 percent versus 7.8 percent).

3.4 Comparing Data Collected by Enumerators Versus the Mail

In comparing the error rates for persons on mail return versus enumerator completed questionnaires, it is clear that self response results in a lower rate of erroneous enumeration (3.1 versus 8.7 percent). Although it may be true that self response provides better quality data, we should not be quick to blame this on the enumerators or assume better training or procedures would remedy this problem. The characteristics of the persons and the housing units that are enumerated in followup activities may be the real source of the problem. If we compare the error rates by type of error (Graph 3.4.1) we note the following:

- There is no significant difference in the rate of geocoding errors over mail return and enumerator completed forms (0.4 percent). This was expected. There is no reason to expect an effect due to method of data collection.
- Fictitious persons were more likely to be found on enumerator completed forms (0.5 versus 0.05 percent). This also should have been expected.
- Duplicates occurred less frequently on mail return (1.0 percent) versus enumerator filled (3.5 percent) forms.
- Other errors occurred at a higher rate on enumerator filled forms (4.3 percent) versus mail return forms (1.7 percent). This type of error had the highest rate over both form types and suggests that it is not always clear to respondents and enumerators who to include as census residents. Enumerators deal more often with complex households where deciding who to include may be much more of a problem. These errors may also be upwardly biased due to potential recall error during PES followup.



4. CONCLUSIONS/RECOMMENDATIONS

Additional analysis is needed to better understand the causes of enumeration errors. Detailed analyses of the demographic characteristics of the erroneously enumerated persons may identify population subgroups at risk. These data suggest that we detect fewer coverage errors on questionnaires that are completed by household members that return their questionnaires by mail. The data also confirm hypotheses that timely data collection is critical to obtaining quality data. This is true for both mail return households and households that were enumerated in followup activities.

The most frequent types of errors appear to be due to residence rule violations (other errors) and duplication. Testing of improved roster questions and instructions on the residence rules is suggested.

Similar analysis of the characteristics of households with missed persons will allow us to determine if the same types of households and housing units are subject to these types of coverage errors.

1. Mary H. Mulry and Bruce D. Spencer, 1990 Decennial Census Preliminary Research and Evaluation Memorandum No. 149, Accuracy of Undercount Estimates for the 1990 Census.

2. Mary M. Mulry, 1990 Preliminary Research and Evaluation Memorandum No. 165, 1990 Post Enumeration Survey Evaluation Project P16 - Total Error in PES Estimates for Evaluation Post Strata.