QUALITY OF DATA KEYING FOR MAJOR OPERATIONS OF THE 1990 CENSUS

Kent Wurdeman, Bureau of the Census
Bureau of the Census, Washington, D.C. 20233

KEY WORDS: Error rate, Cause, Impact

I. INTRODUCTION

Keying was a major component of data collection for the 1990 census. All updates of address lists from field coverage operations were keyed into the master address file, and write-in responses to questions such as race, ancestry, occupation and place of birth were collected from the questionnaires through keying operations.

This paper discusses results from independent studies designed to evaluate two keying operations, race write-in keying and precanvass address list updating. The objectives of the evaluations were to estimate the quality of the keyed data and determine the impact of the keying errors, to determine the causes of error, and to assess the ability of the quality assurance operation to provide accurate quality information for feedback and analysis. These objectives were obtained by producing a file of keyed data independent of the census keying operation.

II. METHODOLOGY OF CENSUS KEYING OPERATIONS

A. Race Write-in

The census questionnaires requested information on race for all persons. Respondents had the option of selecting one of the specific categories listed on the questionnaire or entering a write-in answer in one of two boxes. One box was use to identify an American Indian tribe and the other box was use to identify an Asian/Pacific Islander race or a race not listed.

B. Precanvass

The precanvass operation was performed in urban and major suburban areas to verify the accuracy and completeness of the address list. Census enumerators canvassed streets with address registers, adding addresses missing from the lists, making corrections, and deleting duplicate, nonexistent and commercial addresses. At the end of the field operation, these updates were keyed at four processing offices.

III. EVALUATION METHODOLOGY

The independent keying was performed about a year after the census. For race keying, a one percent sample of questionnaires was selected. Write-in responses to the race question were keyed by two keyers. If there was a difference between what the two keyers keyed, a third person looked at the source documentation and the two keyed entries and determined the proper entry, which was incorporated into the final evaluation file. The methodology for the precanvass keying study was similar.

For the purpose of this study, it is assumed that the keyed responses on the final evaluation file accurately represent the data on the questionnaires/address registers. Conclusions and statements about the quality of the data produced in the census keying operations are made using the evaluation file as the basis for comparison.

Errors were determined by matching each keyed entry in the final census race file to the corresponding entry in the final evaluation file. If the census version did not exactly match the corresponding evaluation version, it was determined to be in error. For race keying, an error was determined to be a critical error if the difference caused by the error was such that the census version would be coded differently than the evaluation version. For precanvass keying, a keying error was determined to be critical if the difference between the census version and the evaluation version was significant enough to

Note: This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the author and do not necessarily reflect those of the Census Bureau.

potentially affect the deliverability of a census questionnaire to the address.

IV. LIMITATIONS

A. Determination of Critical Errors

For race write-in keying, a keying error was determined to be critical if the difference between the census entry and the evaluation entry was such that the two entries would be coded differently. Therefore, the code that would be assigned to an entry had to be determined in order to classify an error as critical, and this determination of code assignment was made by the analyst for this evaluation. The analyst is not a race expert and since the assignment of codes is sometimes a subjective decision, there may be instances where the correct or most appropriate code assignment was not determined.

For precanvass keying, a keying error was determined to be critical if the difference between the census entry and the evaluation entry was significant enough to potentially affect the deliverability of a census questionnaire to the address. The determination of whether a difference was critical was made by the analyst for this evaluation. This determination was somewhat subjective.

B. Determination of Causes of Error

Part of the results involves a discussion of the causes of error. Causes were determined by comparing the keyed entries to the source documentation, i.e. the microfilm of questionnaires for race write-in keying and address registers for precanvass keying. In some cases categorizing the errors into causes depended on the judgement, or educated guess, of the analyst performing this evaluation.

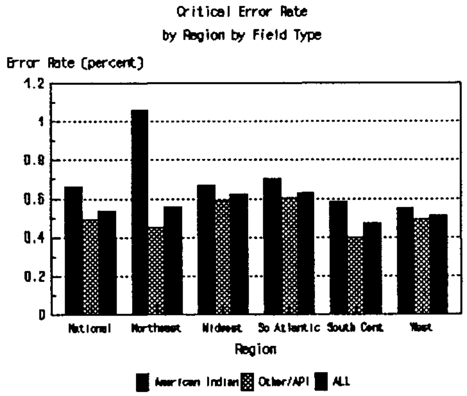
V. RESULTS

Race Keying

A. Quality of the Final Census Race File

Overall, the quality of the race keying was very good. Figure 1 shows the percentage of critical errors in the final census race file by field type and by region. At the national level, the overall estimated field error rate is 0.54 percent. The estimated field error rate for the American Indian field is 0.66 percent, which is significantly higher than the estimated field error rate for the other/API field, 0.49 percent. The field error rate for the American Indian field is higher than the error rate for the other/API field across the country. At the region level, the field error rate for the American Indian field ranges from 1.06 percent for the northeast to 0.55 percent for the west, and the field error rate for the other/API field ranges from 0.7 percent for the south atlantic to 0.4 percent for the south central.

Figure 1 - Quality of the Final Census Race File

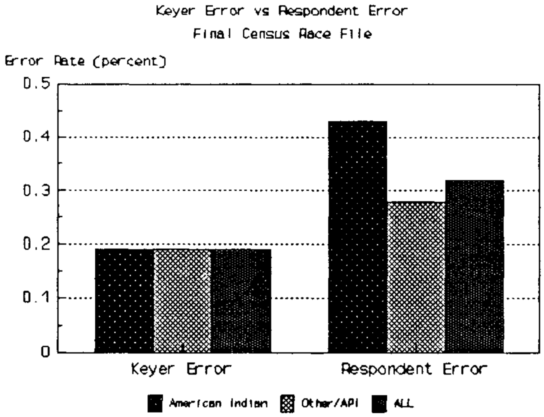


1. Type of Critical Errors: Keyer and Respondent

All errors were classified as one of two types, keyer error or respondent error. Keyer errors are a result of the keyer making some procedural or keystroke mistake. Respondent errors occurred when the data entered on a questionnaire took some form which the keying procedures did not address, sometimes resulting in a difference in interpretation between the census keyer and an evaluation keyer. Examples of respondent errors are writing outside of the box, crossing out a write-in response, answering with a "none" or "same", an illegible response, etc.

Figure 2 shows the percentage of keyer errors and respondent errors in the final census race file by field type at the national level. The estimated total keyer field error rate is 0.19 percent, the estimated keyer field error rate for the American Indian field is 0.19 percent, and the estimated keyer field error rate for the other/API field is 0.19 percent. The estimated total respondent field error rate is 0.32 percent. The estimated respondent field error rate for the American Indian field is 0.43 percent which is significantly different from the estimated rate for the other/API field, 0.28 percent. Approximately 62 percent of the critical errors can be classified as respondent errors.

Figure 2 - Critical Error Rate



2. Cause of Keyer Error

After determining which fields were in error, the source documentation from which the responses were keyed was inspected in order to determine the cause of the error. Keyer errors were categorized into the three following causes:

a. **Wrong Column/Field** (65.8 percent) - The census keyer/verifier keyed a response that was entered in a different (usually adjacent) column or in a different write-in box in the proper column.

b. **Corrected/modified** (19 percent) - The census keyer/verifier intentionally corrected a response that was misspelled or modified a response in some form that was deemed to be more appropriate.

c. **Other Nonsubjective** (15.2 percent) - The census keyer/verifier failed to correctly key a legible response and the cause for the error is not readily apparent.

3. Respondent Error

After determining which fields were in error, the source documentation from which the responses were keyed was inspected in order to determine the cause of the error. Respondent errors were categorized into the six following causes:

a. **Subjective** (8.4 percent) - The write-in response is very difficult to read or is partially illegible and the interpretation differed between the census keyer/verifier and the evaluation keyer/reviewer.

b. **Erased** (26.7 percent) - The write-in response appears to have been erased but the response is still faintly legible. The census keyer/verifier ignored the response, i.e. keyed it as a blank, while the evaluation keyer/reviewer keyed the response, or vice versa.

c. **Outside box** (9.9 percent) - The response includes more than one word and a portion of the response is written outside of the write-in box. The census keyer/verifier keys the portion of the response falling outside of the box while the evaluation keyer/reviewer ignores that portion, or vice versa.

d. **Crossed out** (32.8 percent) - The response appears to have been crossed out but is still legible. The census keyer/verifier ignored the response, i.e. keyed it as a blank, while the evaluation keyer/reviewer keyed the response, or vice versa.

e. None/na/same (8.4 percent) - The response is some uncodable entry such as "none" or "N/A" which is ignored by the census keyer/verifier but is keyed by the evaluation keyer/reviewer, or vice versa. If the response is "same", the census keyer/verifier keyed the response which "same" refers to while the evaluation keyer/verifier keyed "same", or vice versa.

f. Other (13.7 percent) - The error does not fall into any of the above categories.

In most cases of respondent error, the census keyer attempted to judge the respondent's intentions and modified the entry accordingly, whereas the evaluation keyer, having perhaps received more explicit instructions to key exactly what was entered in the write-in box, accurately reflected the actual response rather than the intended answer. Therefore, it could be reasoned that the census version, while being assessed a critical error, may in fact contribute less error to the race tabulations than the evaluation version would have contributed.

Precanvass Results

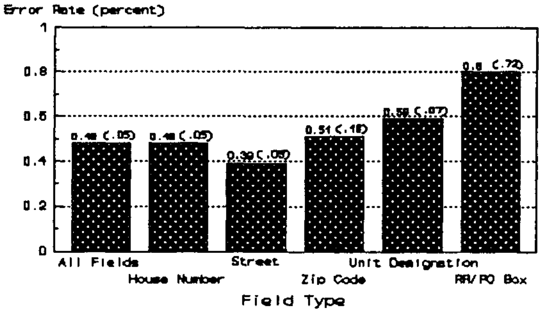
A. Adds

Addresses which were missing from the address registers were added on pages reserved just for adds, and all of the fields on each line were keyed. Each address line contained fields for geocode information and address information. This report will focus on fields relating to address information necessary for mailing, i.e. house number, street name, zip code, unit designation, and rural route/PO box number.

1. Quality of the Final Precanvass File

In this operation also, the keying quality was very good. Figure 3 shows the percentage of critical errors in the final precanvass file by field type for add cases. The overall estimated field error rate is 0.48 percent. The rural route/PO box field has the highest field error rate, but this field occurred infrequently in precanvass areas and the errors were clustered in a few areas resulting in a high standard error.

Figure 3 - Critical Error Rate
by Field Type
For Add Cases - All Mailing Address Fields

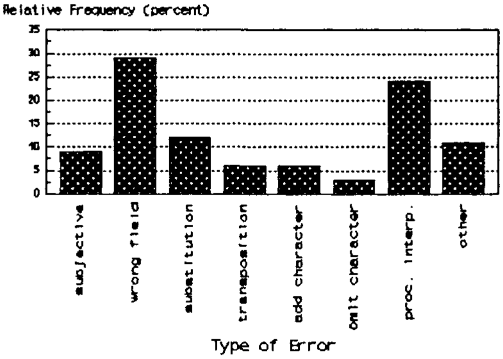


2. Cause of Error

Figure 4 shows the distribution of errors by cause of error for the mailing address fields. About 9 percent of the critical errors were subjective errors. Like race, an error was categorized as subjective when information on the address register was difficult to read and the interpretation differed between the census keyer and the evaluation keyer. Subjective errors usually occurred in the house number and unit designation fields.

Figure 4 - Distribution of Errors by Type

For Add Cases - All Mailing Address Fields



About 24 percent of the errors were caused by a difference in procedural interpretation. A difference in procedural interpretation arose when the information on an address line was in some form which the procedures did not explicitly address, requiring some judgement for resolution, and the census keyer and the evaluation keyer handled the situation differently. About 65 percent of the street name errors were differences in procedural interpretation.

About 29 percent of the critical errors were a result of the census keyer entering information from the wrong field on the page, usually from an adjacent line or column.

In the analysis of the race keying discussed above, a distinction was made between keyer error and respondent error. The same distinction can apply to the precanvass keying results. The causes of error such as wrong field, substitution, transposition, and adding or omitting a character are the result of a keyer making some procedural or keystroke mistake and would be classified as keyer errors. Subjective errors and differences in procedural interpretation would be given a classification similar to respondent error.

3. Impact of Errors

For a particular address line that is keyed, there is a version keyed during the census production and a version keyed independent of the census for the purpose of this study. For cases with a critical keying error, the census version was added to the address list during the precanvass keying operation, and the evaluation version, which is assumed to be the correct version, was not added. This could have had an adverse impact in two ways with the most serious consequence being that the residence represented by the evaluation version was never added to the address list and was not captured in the census. A lesser consequence resulted in the evaluation version being added to the address list in some coverage operation following the precanvass operation. It was also possible that the census version provided enough information to represent the same residence as the evaluation version, resulting in no loss of coverage and no extra burden on subsequent coverage operations.

Table 1 shows the distribution of critical errors by outcome and by mailing address field. Nearly 75 percent of the critical errors which resulted in a missed housing unit occurred in the house number field, yet most house number keying errors still had no major impact on census coverage. Over 75 percent of the critical errors which necessitated that a housing unit be added during a subsequent coverage operation occurred in the house number and unit designation fields. About half of the critical keying errors in the mailing address fields, as determined for this study, had no impact on coverage.

Most of the critical errors in the zip code field were due to a difference in procedural interpretation, because the field was left blank in the address register and the census keyer keyed a zip code based on information from other address lines while the evaluation keyer keyed the field as a blank. This explains why 87.5 percent of the

Table 1 - Impact of Critical Errors
Distribution of Errors by Outcome by Mailing Address Field

Outcome	Relative Frequency (percent)				
	Mailing Address	House Number	Street	Zip Code	Unit Desig
1. Housing unit was not captured in the census	15.3	36.1	7.2	0.6	11.9
2. Housing unit was added in a subsequent coverage operation	30.7	37.5	16.4	4.8	30.5
3. No impact - housing unit was captured	31.9	22.2	10.9	7.1	47.4
4. Inconclusive - evaluation version is incomplete	21.5	4.2	63.7	87.5	11.9

cases with a zip code critical error have an incomplete address for the evaluation version. For similar reasons, 63.7 percent of the cases with a critical error in the street name field have an incomplete address for the evaluation version.

In absolute terms, approximately 14,000 housing units were not captured in the census and approximately 21,000 housing units were added in subsequent coverage operations as a result of keying errors in add cases. Considering that almost 6 million housing units were added during the precanvass operation, the impact of keying errors on census coverage was very small.

B. Corrections

During the precanvass operation, enumerators could make corrections to addresses. Any of the mailing address fields could be corrected except for the house number. If a correction was made, only the particular field corrected was keyed.

1. Critical Error Rates

The critical error rate for correction cases, for the mailing address fields, was 0.79 percent. Most of the corrections were made to the unit designation.

2. Distribution of Errors

About 43 percent of the errors were due to subjective differences caused by corrections which were difficult to decipher. About 21 percent were due to keying an entry from the wrong field. Often the unit designation field and unit description field were mixed up. Nine percent were keystroke substitution errors, and about 23 percent of the errors were a result of a correction not being keyed.

3. Impact of Errors

About 25 percent of the cases with a critical error resulted in a housing unit not being captured in the census. These cases consisted mostly of a correction to the unit designation. In absolute terms, approximately 6000 housing units were not captured in the census as a result of keying errors in correction cases. Considering that almost 3 million addresses were corrected during the precanvass operation, the impact of these keying errors on census coverage was very small.

VI. CONCLUSION

A large proportion of the critical errors on the final census race file and on the final precanvass file, particularly errors in the street name and zip code fields, are due to differences in procedural interpretation which occurred when the information entered by a respondent (or enumerator) was in some form which the procedures did not explicitly address, requiring some keyer judgement for resolution. Procedures for future keying operations should explicitly address these situations so that the keying of these cases will most accurately reflect the intentions of the respondent (or enumerator) and minimize the amount of keyer judgement involved.

Although it is difficult to precisely measure the impact of critical errors, after examining the final census status for the census version and evaluation version of cases with a critical error, it appears likely that the critical errors did place additional burden on coverage operations that followed precanvass and that some relatively small number of housing units were not captured in the census as a result of keying error.