Quality Control of Data Entry for the American Community Survey and the Impact of Errors on Data Quality[1]

Andre Williams, Dale Garrett and Rita Petroni
Andre Williams, U.S. Bureau of the Census, Washington, D.C. 20233
andre.l.williams@census.gov

Key Words: mailout, keying, data capture, consequential errors, quality control

Abstract

The American Community Survey (ACS) is being developed by the U.S. Census Bureau to replace the decennial census long form sample. The ACS collects important socioeconomic data in three-month cycles. The ACS mails out questionnaires during month one, then follows up with nonrespondents with computer assisted telephone and personal interviewing in months two and three, respectively. The paper questionnaires from mailout are keyed. The ACS has a quality control program which samples keyed documents and estimates keying error rates. A sample of errors from the 2002 ACS was examined to measure the impact of keying errors on data quality. Examination of these errors led to the classification of some errors as inconsequential since they were not expected to affect data quality. This paper estimates ACS keying error rates, provides an estimate of the proportion of errors that affect data quality, compares ACS error rates to error rates of data capture methods used by Census 2000, and describes the distribution of the main types of keying errors.

1. Introduction

The American Community Survey (ACS) is being developed by the U.S. Census Bureau to provide the data traditionally collected decennially by the census long form sample. In the near future, the ACS sample will be large enough to replace the census long form sample with a five-year rolling sample.

The ACS collects data through three modes in a three-month cycle.

- During the first month, the ACS collects data through mailout/mailback.
- During the second month, any mail nonrespondents for which telephone numbers are available are sent to one of three telephone centers for Computer Assisted Telephone Interviewing (CATI).
- During the third month, a subsample of cases which could not be interviewed by mail or by CATI are sent to field operations for Computer Assisted Personal Interviewing (CAPI).

Keying is done only for the mailout/mailback portion which comprises more than half of the total number of completed ACS interviews.

It is important to the ACS to maintain the integrity of the data collection process by controlling potential sources of error. The keying quality control program is designed to minimize keying errors that occur during the keying of the ACS questionnaires.

This paper begins with a discussion of keying, Quality Control (QC) and the methodology used in calculating error rate estimates. The results section includes the estimates of error rates, comparison of the estimated outgoing error rates from ACS and Census 2000, and the distribution of the main types of keying errors. We also look at the impact of data entry errors on data quality. The last section provides conclusions based on the analysis.

2. Methodology

2.1 Keying and Keying QC

Three main tasks are performed during the keying operation: keying, verification, and review. See the flowchart in the Appendix for the keying and keying QC process.

Keying is done in batches (work units) which have a maximum of fifty questionnaires each.

After batches have been keyed, the verifier (a second person), verifies the original keyer's work unit. Verification requires the verifier to key a systematic sample of questionnaires from the previously keyed batch.

[1]This report is released to inform interested parties of research and to encourage discussion. The views expressed on operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

A third person (usually the lead keying operator) reviews the discrepancies between the keyed and verified questionnaires and assigns error codes to each discrepancy. Error codes indicate whether the discrepancy is an error by the original keyer or the verifier. Even though many of the fields with errors are in write-in fields which may contain multiple errors, only one error is counted for the field. A count of the number of errors and an error rate are assigned to each work unit.

At the completion of error classification, all fields with errors by the original keyer are replaced with the verifier's field entry. In rare cases where both the original keyer and verifier are in error, the error will be recycled back to the keying supervisor for correction.

Through training and verification, the keying QC plan reduces keying errors introduced during the data entry of mailback questionnaires. A sample of at least 4 percent of the questionnaires from batches are verified as part of the keying QC process.

A keyer progresses through 3 stages: training, prequalification, and qualification.

During the training stage, a keyer is considered a trainee and must key a batch with less than or equal to a 1.5 percent error rate to advance to the pre-qualified stage.

After advancing to the pre-qualified stage, a keyer is considered a pre-qualified keyer but must key another batch with less than or equal to a 1.5 percent error rate to advance to the qualified stage.

After advancing to the qualified stage, a keyer is considered a qualified keyer but is expected to continue keying each batch with less than or equal to a 1.5 percent error rate.

The batch error rate is the error rate from all the questionnaires from the keyed batches of trainees' and pre-qualified keyers. For the qualified keyers, the batch error rate is based on a sample of questionnaires from the keyed batches.

To ensure that keying quality is maintained, any batches with error rates that exceed 1.5 percent are rejected and returned to the original keyer for batch repair. During batch repair, the rejected batches are completely rekeyed and corrected by the original keyer. The keyer keys over the original keyed fields in batch repair. When the original keyer keys in something different than what was keyed the first time, the system beeps and the original keyer has the opportunity to rekey the correct entry or accept the entry keyed. If the original keyer repeats the same error that was found during verification the first time, the system will not beep and the same error will be found if in sample during the second verification. Once the rejected batches have been rekeyed and corrected, they are sent back to verification to be independently sampled and verified (Grice, 2001; Corby, 2001).

2.2 Types of Error Rates

When evaluating the quality of keying, it is important to examine the error rates obtained as a result of the errors made during keying. The three error rates discussed in this paper are: incoming, outgoing, and consequential. These error rates are based on two reports obtained from the keying QA staff of the National Processing Center (NPC). These reports cover two periods. The June 2, 2001 report includes summaries of data sampled from documents keyed during September 2000 through May 2001. The January 4, 2003 report includes summaries of data sampled from documents keyed during September 2001 through December 2002. We now discuss the calculation of these error rates.

2.2.1 Incoming and Outgoing Error Rates

The two error rate estimates discussed here are incoming and outgoing. The total number of incoming errors from each keyer's work including trainees, pre-qualified, and qualified keyers are used to obtain estimates for the incoming error rate. The incoming errors are defined as the errors which may be identified in batches during the verification process prior to batch repair. The total incoming error rate was derived by dividing the estimate of total incoming errors by the total number of keyed fields.

Outgoing errors are defined as the errors that remain in batches after some errors in batches are corrected. Only about 6 percent of all questionnaires are verified. All of the errors found in the sample questionnaires are corrected during verification. The number of outgoing errors was estimated by subtracting an estimate of the number of fields with errors in sample questionnaires that were corrected when the sample was verified from the estimate of total incoming errors. The outgoing error rate was then formed by dividing the estimate of

the outgoing errors by the total number of keyed fields.

The incoming and outgoing error rate estimates were derived from NPC summary report data. These reports supplied the tallies for each group of batches sampled at the same rate. We considered each group to be a stratum. There were 28 strata in the June 2, 2001 report and 29 strata in the January 4, 2003 report.

The total incoming error rate is defined as:

$$p_{inc} = \frac{\sum_{h=1}^{j} (N_h * p_h)}{N}$$

The outgoing error rate is defined as:

$$p_{out} = \sum_{h=1}^{j} \frac{(N_h * p_h - c_h)}{N}$$

where,

h = stratum

N_h = total number of fields in the batches in stratum h

N = total number of fields in all the keyed batches

$p_h = c_h/n_h$ = estimated proportion of fields with errors in the sample from stratum h

c_h = unweighted number of fields with errors found in the sample questionnaires in stratum h. These errors were corrected when the sample was verified. For trainees and pre-qualified keyers, the entire batch or work unit was in sample.

n_h = unweighted number of fields in the sample from stratum h

2.2.2 Consequential Error Rate

Some keying errors likely do not affect survey estimates. We consider these errors inconsequential. These include: the misspelling of a person's name, most typographical errors in a geographical field (such as state, county, and city), some typographical errors in street names, many typographical errors in the name of the employer, etc. Many of the errors in a geographical field (U.S. Census Bureau, 2003) or in a field such as a name of the employer were considered inconsequential since they were likely to be corrected during the geography, industry, and occupation coding.

Here are some examples of inconsequential errors that we expect to not affect ACS estimates or the analysis of data:

Example 1

Field With Discrepancy: Name of Employer

Original keyer keyed: St Louis Univer**is**ty

Verifier keyed: St Louis Univer**si**ty

Reason Inconsequential: ACS will likely code the employer as a university.

Example 2

Field With Discrepancy: Name of state where person was born

Original keyer keyed: Mi**gh**igan

Verifier keyed: Mi**ch**igan

Reason Inconsequential: During the clerical geographic coding phase, 'Mighigan' will be coded as 'Michigan'.

Example 3

Fields With Errors:	Kind of Work	Activity
Keyer keyed:	Clerk Sorting	1
Verifier keyed:	Clerk	Sorting

Additional Relevant Information: A third field indicated that the clerk worked for the U.S. Post Office.

The Error Was: The keyer combined the fields for kind of work and job activities.

Reason Inconsequential: In a later operation, an occupational coder uses all of the related fields so knows that the person was a postal clerk whose main activity was sorting. Because of this, the field, kind of work will be assigned the appropriate coded value.

We modified the estimate of outgoing error rate to obtain a rate that measures the rate of errors that potentially affect the quality of estimates. We call this the **consequential error rate**.

We produced an estimate of the consequential error rate from a sample of 96 discrepancy printouts (Garrett, 2003a). The errors in the sample which could not be shown to be inconsequential were classified as consequential. The proportion of consequential errors in the sample of errors was determined by dividing the unweighted number of consequential errors by the total number of errors in the sample. The estimated consequential error rate was then determined by multiplying the proportion consequential times the outgoing error rate. The variance of the proportion of errors which are consequential was estimated for a clustered simple random sample with variable size clusters (Cochran, 1977).

Errors in questionnaires that did not go through verification have the potential to affect survey estimates or analysis of data. Below are some examples of types of errors in the outgoing product that may affect analysis or estimates:

Example 1

Field With Error: Number of rooms in housing unit (checkbox)

Original keyer keyed: 2 (2 rooms)

Verifier keyed: 1 (1 room)

Reason Consequential: A discrepancy in this field could affect estimates of the number and percent of housing units with 1 room or 2 rooms.

Example 2

Field With Error: Place of Birth (checkbox)

Original keyer keyed: 2 (Born outside the United States)

Verifier keyed: 1 (Born in the United States)

Reason Consequential: A discrepancy in this field could affect estimates of the number and percent of the population born in the United States or outside the United States. However, an error in this field could be inconsequential if the field for the state in which the person was born was filled, which could be used to edit the checkbox.

Example 3

Field With Error: Marital Status (checkbox)

Original keyer keyed: 2 (Widowed)

Verifier keyed: 3 (Divorced)

Reason Consequential: A discrepancy in this field could affect estimates of the number and percent of persons in the population that are widowed or divorced.

3. Results

The analysis for this paper is based on data from two reports. The error rate from a single year could not be determined from the reports, since the tables in the reports did not break down the summaries by month or by year. The data provided were not available at lower levels, such as for each sample document and each batch. The summary data was limited to group of batches sampled at the same rate. All tests for statistical significance were performed using alpha = 0.1.

3.1 Incoming and Outgoing Error Rate Estimates

Table 1 shows the error rates and associated standard errors for the two time periods from the June 2, 2001 and the January 4, 2003 summary reports. Both the incoming and outgoing error rates for each time period were well below 1 percent. Both the incoming and the outgoing error rates appear to have decreased between the two time periods. However, the differences shown were found not to be statistically significant.

Table 1. Data Entry Error Rates for ACS

Error rates	June 2, 2002 Report September 2000 - May 2001	January 4, 2003 Report September 2001 - December 2002
Incoming Error Rate (SE)	0.65% (0.06%)	0.54% (0.05%)
Outgoing Error Rate (SE)	0.60% (0.05%)	0.51% (0.04%)
Number of questionnaires in sample	17,121	18,430
Number of fields in sample	2,449,001	2,840,517
Total Number of fields	43,104,729	57,337,877

3.2 Consequential Error Rate

Table 2 shows the consequential error rate and associated standard errors for the two time periods. The proportion of outgoing errors estimated to be consequential was 0.3984. After applying 0.3984 to the outgoing error rate, consequential error rates were estimated as 0.24 percent and 0.20 percent for the two time periods.

Table 2. Data Entry Consequential Error Rate for ACS

Consequential Error Rate	
September 2000 - May 2001	0.24% (0.05%)
September 2001 - December 2002	0.20% (0.04%)

3.3 Comparison of the ACS Outgoing Error Rate to the Error Rates of Census 2000

Since the ACS is expected to replace the census long form, it would be good to know if the ACS keying error rate is at least as good as the data capture error rate for census long forms. So, we compared the ACS outgoing error rates to the mean error rates for long forms (Conklin, 2003).

We found that the outgoing keying error rate for ACS compares favorably to the error rates of Census 2000. For Census 2000 keying was not done directly from paper questionnaires as was the case for ACS. For the Census a digital image of each questionnaire was first created. Then three methods were used to capture data from images of paper questionnaires: optical character recognition (OCR) for write-in fields, optical mark recognition (OMR) for check boxes and key from image (KFI) for fields with unclear characters or marks. As Table 3 shows, the combined data capture error rate for census long forms was significantly above 1 percent. The data entry error rate for ACS keying was below 1 percent. The difference in the ACS outgoing error rate and the Census 2000 mean error rate is a by-product of the difference in modes of data capture from the paper questionnaire.

Table 3. Estimated Outgoing Error Rates from ACS and Census 2000

ACS Outgoing Error Rate		Census 2000 Data Capture
September 2000 Through May 2001	September 2001 Through December 2002	
0.60% (0.05%)	0.51% (0.04%)	1.80%* (0.02%)

*Source: Conklin, 2003. Error Rate is a mean error rate from long forms only across all three modes of data capture

3.4 Distribution of the Main Types of Keying Errors

To better understand the nature of keying errors in the ACS, the most common kinds of outgoing errors were examined: data omission, finger error, procedure error, and code error. **Data omission** means that the field was left blank even though the questionnaire contained data. However, if the field was left blank because of a procedure error (not skipping to the correct field) then a **procedure error** might be marked instead of a data omission. A **finger error** is a typographical error associated with a write-in field. A **code error** is an error caused by difficult to read handwriting and is counted as an error, but is not

charged against the keyer (Grice, 2001; Klein, 2001 and 2003).

Table 4 shows that code and finger errors account for more than half of the errors. These are generally errors in a name of a person or business, state name, address, job description, etc. These errors from write-in fields are frequently inconsequential as discussed above.

Table 4. Percent of Total Outgoing Errors for the Main Error Types

Type of Errors	September 2000 Through May 2001	September 2001 Through December 2002
Data Omission	18.5	15.6
Procedure Error	14.9	18.9
Finger Error	31.7	33.9
Code Error	34.7	31.3
All Other Error Codes	0.2	0.3
Total Number of Errors	14,101	13,906

4. Conclusions

Based on these findings we believe that the data entry process for ACS introduces very little nonsampling error into ACS data. The keying QC procedures are effective in controlling the error rates. The few keying errors which remain in the outgoing product are usually of little or no consequence to ACS.

5. References

Cochran, W. G., *"Sampling Techniques,"* third edition: New York: John Wiley & Sons, 1977, p.66.

Conklin, J. D., *"Mean Nonblank Error Rates for Data Capture of Long Forms As Computed From Evaluation of the Quality of the Data Capture System and the Impact of the Data Capture Mode on the Data Quality, K.1.b,"* internal Census Bureau memorandum, May 15, 2003.

Corby, C., *"Quality Assurance Summary of ACS Keying"* document number E-1 of the American Community Survey Research and Evaluation Program, July 9, 2001.

Garrett, D. (2003a), *"Detailed Explanation of the Preliminary Consequential Error Rate of Data Entry for the American Community Survey (ACS),"* internal Census Bureau memorandum, October 9, 2003.

Garrett, D. "*Results of the 2000-2002 Keying Quality Assurance Evaluation of the American Community Survey (ACS),*" internal Census Bureau memorandum from Killion to Singh, October 15, 2003.

Grice, M. T., *"American Community Survey - Quality Assurance for Data Entry Operations,"* National Processing Center memorandum number 4200-907-K, January 24, 2001.

Klein, D., *"The American Community Survey – DEC Data Entry Procedure,"* National Processing Center memorandum number 4200-910-K, March 8, 2001.

Klein, D., *"The American Community Survey – DEC Data Entry Procedure,"* National Processing Center memorandum number 5935-305-K, February 25, 2003.

U.S. Census Bureau, "*American Community Survey Operations Plan*", Release 1: March 2003, p. 28.