Proceedings of the Annual Meeting of the American Statistical Association, August 5-9, 2001

CONSISTENCY OF CENSUS 2000 POST-STRATIFICATION VARIABLES

James Farber and Peter Davis, U.S. Bureau of the Census[1]
James Farber, U.S. Bureau of the Census, 4700 Silver Hill Road, Stop 9200, Washington, DC 20233-9200

Key Words: Accuracy and Coverage Evaluation; Categorical Data; Heterogeneity Bias

1. Introduction

The Accuracy and Coverage Evaluation (A.C.E.) consisted of two independent samples. The first was a sample of the population in selected A.C.E. sample areas, known as the Population or P sample. An estimate of the proportion of the population omitted from the census was obtained by matching these people to census records. The second was a sample of the census enumerations in the A.C.E. sample areas, known as the Enumeration or E sample. An estimate of the proportion of correctly enumerated census people was determined using the results of matching the P sample to the census, checking for duplication in the census records, and re-interviewing when needed to determine the inclusion status of each E-sample person record. Together, these proportions estimated the omissions from and erroneous enumerations in the census.

The A.C.E. included dual system estimates for up to 448 post-strata for the 50 states and the District of Columbia (Davis, 2001). The people in the P and E samples were assigned to post-strata based on race, Hispanic origin, age, sex, and tenure (owner or renter), along with some geographic and processing variables. Ideally, a P-sample person record and an E-sample person record that matched have consistent values for the post-stratification variables, but in reality this may not occur. For example, a person reported his race as Native American when he filled out his census form but his roommate said he is White in the A.C.E. interview. If a matched person did not have consistent characteristics in his P-sample and E-sample records, then that person was classified into two different post-strata when estimating the net proportions of people missed or correctly enumerated in the census. Persistent differences in the classification of person records in the census and the A.C.E. may increase heterogeneity bias in the coverage estimates, which rely on the assumption that people in the same post-stratum had the same probability of capture in the census. This type of heterogeneity is also known as classification error.

A number of factors may create inconsistent P- and E-sample data. Imputation of missing data caused much inconsistency because the imputation processes of the census and A.C.E. were different and were conducted independently. In this paper we split out imputed and non-imputed cases to clarify the source of inconsistency. For non-imputed cases, inconsistency arose due to factors such as inconsistent reporting, data collection mode, proxy responses, and the time lag between the census and A.C.E. An analysis of these sources of inconsistency is beyond the scope of the current research.

This paper summarizes the consistency of the A.C.E. demographic post-stratification variables and assesses the potential effects of inconsistency. Similar analyses were done on data from earlier census tests (Petroni, 1996A and 1996B; Salganik, 1999) and Census 2000 (Farber, 2001), but the current research includes several extensions. The data for our analysis are provided by the P-sample and E-sample person records who matched. We cannot directly assess the consistency of non-matched people because they were missing either P- or E-sample data. But we get a sense of the possible inconsistency of non-matches by examining what types of person records were in the non-matched universe. We also test for significant inconsistency using categorical data analysis techniques. Finally, we analyze the effects of inconsistency on the A.C.E. coverage correction factors and estimate the heterogeneity bias caused by inconsistency.

2. Assessing Consistency

We define consistency as agreement between the categories of the P-sample and E-sample post-stratification variables. The data do not have to match exactly, but rather they place a matched person record into the same post-stratum group. For example, a person who reported her age as 28 on her census form and 29 in the A.C.E. interview is consistent because her record is in the Female 18- to 29-year-old group of the age/sex post-stratification variable in both the E sample and P sample.

Because imputation was a primary cause of inconsistency, we show consistency results for all matched person records and also for non-imputed person records. We define a person record as non-imputed only if the data for both the P sample and E sample were not imputed. The consistency of non-imputed person records more accurately reflects the level of mis-reporting between the census and A.C.E.

[1] This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress.

2.1 Overall Levels of Consistency

Tables 1 and 2 on the last two pages of this paper show the consistency of the tenure and race/Hispanic origin domain post-stratification variables. Each table gives results for all matched person records and for non-imputed person records.[2] We do not show the consistency of age/sex due to space constraints. Farber (2001) provides more detailed consistency results, including tables for age/sex and cross-classifications of all post-stratification variables.

Cases in the shaded diagonal cells of the tables were consistent, while off-diagonal cases were inconsistent. Overall, about 3.9 percent of matched person records were inconsistent on race/origin, which dropped to about 3.2 percent by removing imputed records. Tenure has about 4.7 percent inconsistency for all matches and 3.8 percent for non-imputed matches.[3] The inconsistency rate for imputed person records is not shown but is almost 12 percent for race/origin domain. However, because few person records needed imputation, the effect of their high inconsistency rate on the overall rate was small.

Consistency has positive correlation with sample size. The two smallest domains, American Indian off reservations and Native Hawaiian or Pacific Islander, had much higher rates of inconsistency than the other domains. We discuss the possible effects of high inconsistency in these domains below. The American Indian on reservation domain had low inconsistency despite its relatively small sample size. This was likely due to the requirement of physically living on a reservation, a non-subjective situation that reduced confusion and mis-reporting and thus increased consistency.

Davis (2001) gives the algorithm for classifying person records into race/origin domains, including those with multiple race. Regardless of how many races a person reported, every person record went into one and only one race/origin domain based on E-sample data, and one and only one domain based on P-sample data.

2.2 Significance Testing for Consistency

Tables 1 and 2 show matched nominal data that can be tested for significant consistency between the P and E samples. The *kappa* statistic (Agresti, 1990) measures the strength of agreement for this type of data, where the variables have the same classifications. Kappa equals zero when the agreement between the P and E samples equals that expected by chance, and it equals one when there is perfect agreement. Table 3 gives 90-percent confidence intervals for the kappa statistics for each of the three demographic post-stratification variables.

Table 3. 90-Percent Confidence Intervals for Kappa

Variable	All Matches	Non-Imputed Matches
Race/Origin Domain	0.923, 0.926	0.933, 0.936
Age/Sex	0.938, 0.939	0.965, 0.966
Tenure	0.888, 0.890	0.910, 0.912

The P- and E-sample data agree strongly. Matched people generally provided responses on their census forms and in their A.C.E. interviews that placed them into the same levels of the demographic post-stratification variables. As expected, the strength of agreement increases by removing imputed person records because imputation increases inconsistency.

2.3 Possible Inconsistency of Non-Matches

The P sample contains the set of records of people interviewed in the A.C.E., which was conducted independently of the census. The E sample is the set of records of people enumerated in the census in the same sample areas as the P sample. The P and E samples were matched in a multi-phase operation as part of the A.C.E. estimation process. The first phase was a computer match based on characteristics like name and age, followed by a clerical match. Certain non-matched cases were then sent to field follow-up to obtain information about their Census Day residence status, which was used in a final clerical match. Despite these efforts, about 10 percent or more of the person records in the P and E samples did not match (Childers *et al,* 2001).

The data in our research come only from matched person records, but non-matches may also be inconsistent if their data do not represent the truth. It is important to gauge the possible inconsistency of non-matches because an analysis based only on matches may be misleading. The A.C.E. estimation process used all E- and P-sample records, including non-matches, and thus non-matches could have contributed to heterogeneity bias. To get a sense of the inconsistency of non-matches, we examined the types of records in the non-matched universe.

One might think the simple fact that they were not matched means non-matched person records had higher rates of inconsistency. But non-matches occurred for

[2] The non-imputed person records are those that were not imputed in the census or in the A.C.E. for the variable under consideration. For example, in Table 1B, tenure was not imputed in either the census or A.C.E. for the 520,715 person records.

[3] Age/sex has 5.1 percent inconsistency for all matches and 2.9 percent for non-imputed matches (Farber, 2001).

many reasons that do not necessarily correspond to poor data quality. For example, about 53 percent of the P-sample person records needing follow-up ended up as true residents who were census omissions.[4] About 72 percent of the E-sample records needing follow-up were correct enumerations (Childers *et al*, 2001). An assumption of the A.C.E. estimation method is that correct enumerations were as likely to be matched as not matched, which implies that their consistency patterns were equivalent to those of matched person records. Likewise, the P-sample non-matches that were census omissions probably had good data with inconsistency rates similar to matches. The other types of non-matches may have more inconsistency. For example, about seven percent of the E-sample follow-up universe ended up as erroneous census enumerations, including duplicates, records for fictitious people, and records with insufficient information for matching. The latter group had insufficient information due to missing data that needed imputation, which thus increased their inconsistency.

But in general, a majority of the P-sample and E-sample cases that went to follow-up resulted in person records wrongly omitted from or correctly included in the census. These cases likely had data quality similar to the matched person records, and thus their inconsistency rates were also probably similar to those of matches. We believe our analysis based only on matched person records provides a close approximation to the results one would obtain from a study that did include the non-matched records. Such a study would be prohibitively expensive, as it would require extensive field work to find and reinterview non-matched people.

3. Potential Effects of Inconsistency

Dual system estimation, the A.C.E. estimation method, includes the assumption that the probability of correct enumeration in the census is similar for all people in the same post-stratum. That is, members of the same post-stratum have homogeneous capture probabilities. Heterogeneity arises when capture probabilities differ greatly within a single post-stratum, and contributes to bias in the A.C.E. estimates.

An inconsistent person record was placed into two different post-strata: one based on P-sample data and the other based on E-sample data. One of these post-strata likely represented a true classification of the person record. The record was a member of the other post-stratum, the "wrong" post-stratum, only because of classification error. Under the homogeneity assumption, the person's capture probability is equal to the capture probability for only one of those post-strata, the "true" post-stratum. Thus, an inconsistent person record can increase the heterogeneity bias in the post-stratum in which it was mis-classified.

The effect of heterogeneity caused by inconsistency varies depending on the amount of inconsistency and the difference in coverage properties between the two post-strata of an inconsistent person record. Inconsistency between post-strata with similar capture probabilities creates negligible heterogeneity. Similarly, very low rates of inconsistency between two post-strata have little effect.

Also, the dual system estimation method is robust to inconsistency because it is a ratio estimator that involves two rates: the correct enumeration rate and the match rate. A simplified version of the net coverage rate from dual system estimation within a post-stratum is $\frac{CE}{E} \div \frac{M}{P}$, where

CE = the estimate of correct enumerations
E = the estimate of total E-sample person records
M = the estimate of total P-sample matches
P = the estimate of total P-sample person records.

Inconsistency generally affects the numerator and denominator of each term in similar ways, meaning the net effect of inconsistency on the estimate is reduced by some canceling. Though the estimate might be biased, it is likely closer to the truth than an estimate from a post-stratification plan that excluded or collapsed over potentially inconsistent post-strata. Alternatives such as collapsing can increase heterogeneity bias as well.

For example, Table 2 shows high rates of inconsistency for race/Hispanic origin domain 2, American Indian off reservations, with many of the inconsistent person records also in domain 7, Non-Hispanic White or other race. The results in Davis (2001) indicate that the coverage correction factor (CCF) for American Indians off reservations was statistically significantly higher than that of Non-Hispanic Whites. Thus the inconsistency in the American Indian off reservations domain likely reduced its CCF. However, because about 70 percent of the matches in the domain were consistent, its CCF was closer to the truth than a factor provided by, for example, collapsing the American Indian off reservations domain with another domain. The effects of this classification error on the White domain were negligible. The large number of consistent person records in that domain made its CCF robust to the small number of American Indians off reservations switching into the White domain.

To quantify the potential effects of inconsistency, we used a method based on Mulry and Spencer (2001). The general idea is to simulate the CCFs under perfect consistency. We assumed the E-sample data were correct because the CCFs are ultimately applied to the census, the source of the E sample. Thus we changed inconsistent P-sample responses in our simulation to agree with the E sample, except when the E-sample item was imputed and

[4] Nearly all of the P-sample and E-sample person records needing follow-up were non-matches.

the P sample was not. In that case, we changed the E sample to agree with the P sample. For simplicity, we did not change any post-strata for non-matches, a deviation from the Mulry and Spencer. We also did not recompute variances, but instead assumed the simulated CCFs have the same variance/covariance matrix as the original CCFs. In the simulation, we considered a person record inconsistent if the P sample differed from the E sample on tenure, age/sex, or race/Hispanic origin domain. About ten percent of the P sample and three percent of the E sample changed post-strata in the simulation.

Table 5 shows the absolute value of the differences between the original CCFs and the simulated CCFs for the race/Hispanic origin domains.

Table 5. Differences between Official and Simulated Coverage Correction Factors by Race/Origin Domain

Domain	Absolute Value of Difference
American Indian on reservations	0.0003
American Indian off reservations	0.0130
Hispanic	0.0005
Non-Hispanic Black	0.0022
Native Hawaiian or Pacific Isl.	0.0112
Non-Hispanic Asian	0.0049
Non-Hispanic White or Other	0.0008

None of the differences are significantly different from zero. As expected, the largest changes occur in the CCFs for the American Indian off reservation domain and the Native Hawaiian or Pacific Islander domain, which had the highest inconsistency rates. Also as expected, inconsistency affects the numerator and denominator of the dual system estimator relatively equally, leading to canceling that reduces change in the CCFs. For example, in the American Indian off reservation domain, the number of correct enumerations dropped about 2.6 percent when consistency was improved in the simulation. But the number of E-sample person records fell about 2.5 percent. So the correct enumeration rate, $\frac{CE}{E}$, did not change much in the simulation. The match rate, $\frac{M}{P}$, was affected similarly.

A further analysis of our simulation showed that none of the 448 post-strata CCFs were changed significantly by correcting for inconsistency. Likewise, reversing our simulation methodology by changing the E sample to be consistent with the P sample did not create any significant differences in the CCFs. Given these results, the non-matches excluded from the earlier assessments of consistency would have to be extremely inconsistent to significantly increase the heterogeneity bias of the CCFs.

4. Conclusions

Inconsistency between the P and E samples was inevitable due to many reasons, such as imputation of missing data and misreporting due to proxy respondents, recall bias, or enumerator or interviewer error. Our analysis suggests the levels of inconsistency observed in Census 2000 and the A.C.E. were not significant. The CCFs were not changed significantly when we corrected the inconsistent data in our simulation. Given the number of matched person records in our data, it seems unlikely that the non-matches would have inconsistency rates great enough to introduce noticeable heterogeneity. Inconsistency appears not to have contributed a significant amount of heterogeneity bias to the A.C.E. estimates.

5. References

Agresti, A. (1990), *Categorical Data Analysis*, New York: John Wiley and Sons.

Childers, D., Byrne, R., Adams, T. and Feldpausch, R. (2001), “Accuracy and Coverage Evaluation: Person Matching and Follow-up Results,” DSSD Census 2000 Procedures and Operations Memorandum Series B-6*, Washington: Bureau of the Census.

Davis, P. (2001), “Accuracy and Coverage Evaluation: Dual System Estimation Results,” DSSD Census 2000 Procedures and Operations Memorandum Series B-9*, Washington: Bureau of the Census.

Farber, J. (2001), “Accuracy and Coverage Evaluation: Consistency of Post-Stratification Variables,” DSSD Census 2000 Procedures and Operations Memorandum Series B-10*, Washington: Bureau of the Census.

Mulry, M., and Spencer, B. (2001), “Overview of Total Error Modeling and Loss Function Analysis,” DSSD Census 2000 Procedures and Operations Memorandum Series B-19*, Washington: Bureau of the Census.

Petroni, R. (1996A), “Disagreement of Characteristics Between R-Sample and Census Linked Cases for Oakland-Preliminary Findings,” Internal Bureau of the Census Memorandum.

Petroni, R. (1996B), “Disagreement of Characteristics Between R-Sample and Census Linked Cases for Oakland-More Findings,” Internal Bureau of the Census Memorandum.

Salganik, M. (1999), “Accuracy and Coverage Evaluation Survey: Consistency of Potential Poststratification Variables,” DSSD Census 2000 Procedures and Operations Memorandum Series Q-10, Washington: Bureau of the Census.

Table 1A. Consistency of Tenure for All Matched Person Records

		E Sample (Census)		Total	% Inconsistent
		Owner	Non-Owner		
P Sample (A.C.E.)	Owner	370,258	11,652	381,910	3.05%
	Non-Owner	14,163	153,572	167,735	8.44%
	Total	384,421	165,224	549,645	
	% Inconsistent	3.68%	7.05%		4.70%

Table 1B. Consistency of Tenure for Non-Imputed Matched Person Records

		E Sample (Census)		Total	% Inconsistent
		Owner	Non-Owner		
P Sample (A.C.E.)	Owner	354,245	8,684	362,929	2.39%
	Non-Owner	10,845	146,941	157,786	6.87%
	Total	365,090	155,625	520,715	
	% Inconsistent	2.97%	5.58%		3.75%

Table 2A. Consistency of Race/Hispanic Origin Domain for All Matched Person Records

		E Sample (Census)							Total	% Incon.
		AI on res.[5]	AI off res.	Hisp.	Black	NHPI	Asian	White		
P Sample (A.C.E.)	AI on res.	11,009	0	34	12	0	0	118	11,173	1.47%
	AI off res.	0	2,223	59	104	0	30	793	3,209	30.73%
	Hispanic	44	136	67,985	610	42	267	4,004	73,088	6.98%
	Black	10	119	496	65,679	6	118	1,423	67,851	3.20%
	NHPI	0	3	31	19	1,671	204	177	2,105	20.62%
	Asian	1	31	107	102	143	19,679	1,062	21,125	6.84%
	White	107	944	5,041	2,589	183	2,105	360,125	371,094	2.96%
	Total	11,171	3,456	73,753	69,115	2,045	22,403	367,702	549,645	
	% Incon.	1.45%	35.68%	7.82%	4.97%	18.29%	12.16%	2.06%		3.87%

Table 2B. Consistency of Race/Hispanic Origin Domain for Non-Imputed Matched Person Records

		E Sample (Census)							Total	% Incon.
		AI on res.	AI off res.	Hisp.	Black	NHPI	Asian	White		
P Sample (A.C.E.)	AI on res.	10,485	0	24	10	0	0	103	10,622	1.29%
	AI off res.	0	2,033	48	84	0	25	706	2,896	29.80%
	Hispanic	28	84	54,116	401	34	177	2,997	57,837	6.43%
	Black	10	94	349	59,441	5	80	1,068	61,047	2.63%
	NHPI	0	3	15	16	1,552	178	147	1,911	18.79%
	Asian	1	15	72	69	110	18,040	740	19,047	5.29%
	White	93	848	3,520	2,073	141	1,722	343,632	352,029	2.39%
	Total	10,617	3,077	58,144	62,094	1,842	20,222	349,393	505,389	
	% Incon.	1.24%	33.93%	6.93%	4.27%	15.74%	10.79%	1.65%		3.18%

[5] The race/Hispanic origin domains are:

- American Indian or Alaska Native on reservations
- American Indian or Alaska Native off reservations
- Hispanic
- Non-Hispanic Black
- Native Hawaiian or Pacific Islander
- Non-Hispanic Asian
- Non-Hispanic White or Other Race