

## How Consistent is Race Reporting Between the Census and the Census 2000 Supplementary Survey?

David A. Raglin and Theresa F. Leslie, United States Bureau of the Census

David Raglin, Bureau of the Census, 4700 Silver Hill Road, Washington, DC 20233-9200

**Key words:** Mode, CATI, CAPI, questionnaire design

### QUESTIONS TO BE ANSWERED

This paper uses the results of a match of people collected in the Census 2000 and the Census 2000 Supplementary Survey (C2SS) to help answer the following questions:

- How consistent is race reporting between Census 2000 and the C2SS?
- What subgroups have inconsistencies in race reporting between Census 2000 and the C2SS?
- What factors, such as data collection mode, could be related to differences in the consistency of race reporting between Census 2000 and C2SS?

### BACKGROUND

#### The American Community Survey and C2SS

The American Community Survey (ACS) is a new survey being tested by the Census Bureau. It is designed to go into production in 2003 as a replacement for the Decennial Census long form, with data collected monthly. The ACS asks questions on education, employment, income, housing value, rent, and many other topics.

Data are collected using three modes—first by mail, then by computer-assisted telephone interviewing (CATI), and then for a subsample of remaining housing units, by computer-assisted personal interviewing (CAPI) using experienced Census Bureau interviewers.

The C2SS was a survey of about 700,000 housing units using ACS data collection methods. It was designed to test how feasible it was to collect long form data at the same time as a Census. The C2SS was the first large-scale national collection of data using the ACS process.

#### Differences in Data Collection Operations

Although Census 2000 and the C2SS were both large survey operations by the same agency, there were differences in how the data were collected.

Census 2000 data were collected using two modes: self-response, and personal interview using a paper-and-pencil instrument. About five in six Census households received the short form, which asked for name, relationship to the first person, sex, age, date of birth,

This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau Publications. This report is released to inform interested parties of research and to encourage discussion.

Hispanic origin, and race—in that order for each person. (These are referred to as the 100% demographic items.) The other one in six households received the long form, which asked the previous items plus many others. By design, C2SS households did not receive Census 2000 long forms—just short forms.

C2SS data were collected using three modes: mail, CATI, and CAPI. All of the forms contained essentially the same content as the Census long form.

There were a few differences in the layout of the race question on the mail forms. The Census 2000 questionnaire asked about each person one by one, using separate pages for each person. To contrast, the C2SS collected the 100% items using a grid format similar to the 1990 Census, with names listed down the side and questions across the top.

However, there were more differences in the data collection methods between Census 2000 and the C2SS used for the personal visit data collection than for the mail data collection. Table 1 compares Census 2000 and the C2SS data collection methods in the field:

*Table 1: Data Collection Differences Affecting Race On Personal Visit Instruments, Census 2000 vs C2SS*

#### Interviewer Questionnaire Medium--

- Census: Paper form, with the race question in a grid like the C2SS self-enumeration form.
- C2SS: A CAPI instrument, which asked all of the 100% items, like race, first for each person (like the C2SS self-enumeration form.)

#### Interviewers--

- Census: Temporary interviewers hired for Census
- C2SS: Permanent Bureau interviewers-- also conduct surveys like the Current Population Survey.

#### Question--

- Census: "Which race does (this) person consider himself/herself to be?"
- C2SS: "Please choose one or more categories that best indicate (person's) race."

#### Comparison of Race: Census 2000 vs C2SS

Race is one of the most high-profile variables collected by the Census Bureau. Since both Census 2000 and C2SS provide race data for roughly the same time period, it was logical to compare the race results from Census 2000 and C2SS. Census 2000 and C2SS collected race using similar rules:

- Race and Hispanic origin were separate questions with Hispanic origin asked before race; "Hispanic"

- is not considered a race.
- There were 15 race categories, including “Some Other Race”. If that was selected, the respondent was to provide a specific race.
- If the respondent answered “American Indian or Alaska Native”, “Other Asian” or “Other Native Hawaiian/Pacific Islander”, he or she was asked to provide a specific tribe or race.

Table 2 is a comparison of the racial distribution for the C2SS as compared to the Census household population of about 274 million people. Since C2SS did not include group quarters, the Census population used in this comparison does not include the group quarters population either.

Table 2: Race Distribution, Census 2000 vs. C2SS

Race	Census %	C2SS %
White alone	75.33	*77.48
Black alone	12.04	*11.75
Asian alone	3.67	3.81
NHOPI alone	0.14	*0.16
AIAN alone	0.88	*0.77
Some Other Race alone	5.50	*3.90
Two or more races	2.45	*2.12

\* C2SS estimate significantly different from Census ( $\alpha = .10$ )  
 NHOPI: Native Hawaiian or Other Pacific Islander  
 AIAN: American Indian/Alaska Native

The largest differences were for White alone-- the C2SS estimate was 2.15 percentage points higher than the Census, and Some Other Race alone--the C2SS estimate was 1.60 percentage points lower.

Table 3 shows this trend was especially strong for people of Hispanic origin. This table shows the percentage of the 274 million residents of households that fall into the given category.

Table 3: Race of Hispanics, Census 2000 vs. C2SS

Race	Census %	C2SS %
All Hispanics	12.64	12.55
White alone	6.05	*7.89
Black alone	0.25	*0.20
Asian alone	0.04	0.04
NHOPI alone	0.02	0.02
AIAN alone	0.15	*0.11
Some Other Race alone	5.34	*3.69
Two or more races	0.80	*0.60

\* C2SS estimate significantly different from Census ( $\alpha = .10$ )

The important thing to note from Table 3 is that the C2SS had a higher proportion of White Hispanics than did the Census by 1.84 percentage points, while the C2SS had a 1.65 percentage points fewer Hispanics of Some Other Race than did the Census. Comparing those differences from the differences for the same items from Table 2, it seems that much of the difference seen in White and Some Other Race in Table 2 was due to Hispanics.

Why are there such differences in the reporting of race by Hispanics? Part of that could be because many Hispanics consider “Hispanic” to be their race (Bates, Martin, DeMaio, and de la Puente, 1995, and Martin, DeMaio, and Campanelli, 1990), while the Federal Government considers Hispanic to be an ethnicity, distinct from race. Therefore, many Hispanic people report their race as Some Other Race. This is true even when the Hispanic origin question is asked before the race question, so that Hispanics have already had the opportunity to report their Hispanic heritage (Bates, Martin, DeMaio, and de la Puente, 1995). However, the same paper says that asking Hispanic origin before race does minimize that effect, as Census 2000 and C2SS both did.

**METHODOLOGY**

Tables 2 and 3 showed differences in the race distribution in the C2SS compared to Census 2000. One method to determine why those differences might be occurring is to compare the responses to the race question for the same people, and if there are differences, to try to understand the reasons for those differences.

**Data Preparation Methodology**

The C2SS data were collected in a series of monthly samples, or panels. Each monthly sample was a nationally representative sample of all housing units in the United States. For example, for the April 2000 panel, questionnaires were mailed in April, followed up with CATI in May, and then in June with CAPI for a subsample of the remaining nonrespondents.

The 360,855 eligible people collected by the C2SS in the March, April, and May, 2000 panels were included in this study, hereafter referred to as the C2SS people. Their C2SS data were collected between March and July, 2000, consistent with the Census 2000 data collection time frame.

The C2SS people were matched to the people in the Census 2000 files. Since this is a study of the consistency of reporting and not a coverage study, the methodology erred on the side of being conservative in what was considered a match. That ensures that almost all, if not all, of the matches are in actuality the same people. However, if reviewed individually, probably more people would be considered a match.

The matching was done in two stages using AutoMatch, a commercial matching software program:

- The C2SS people were matched to Census 2000 people within the same housing unit, using the Master Address File (MAF) ID to identify the housing units, and matching on name, age/date of birth, and sex.
- The remaining unmatched C2SS people were matched to Census people within state (two-digit zip for CA, NY, TX, and FL) using the same variables plus address, and not allowing new matches among people within the same MAF ID. The cutoffs for determining a match were much more stringent for this stage than for the first stage.

The result was that 324,465 of the C2SS people matched Census people, a 90 percent match rate. There were 313,030 C2SS people that matched Census people in the same MAF ID (96 percent of the matches), with the other 11,435 people matched within the state or two-digit zip code.

The data in Tables 2 and 3 represent the final C2SS and household Census 2000 data, after the data are edited, imputed for item nonresponse, adjusted for unit nonresponse, and for the C2SS data, benchmarked to Census 2000 data.

However, this study examined differences in race reporting at the person level, and therefore, we only wanted to use the race information that was actually reported by the respondent. This study used the combination of checkbox and write-in race data provided by the respondents to determine the reported race.

The C2SS was a sample survey, as opposed to the Census, where everyone in the country was asked to provide their race. This study used the weights for C2SS that took into account all stages of sampling, including the additional subsampling for households who were interviewed in C2SS using CAPI. No weight adjustments for unit nonresponse or benchmarking to the Census were used. Only the sampling weights were used because this study used the race responses straight from the respondents without editing or nonresponse imputation.

### Statistical Methodology

The analysis for this paper relies mainly on a few statistics often used when comparing two sets of matched categorical data that are measuring the same phenomena.

The first statistic is the Net Difference Rate (NDR). The NDR is defined as the expected difference between the estimates of the percent in each race group between the Census and the C2SS. For the  $i^{\text{th}}$  race group, it is estimated using the following formula:  $(n_i - n_{i,c}) / n_{i,c}$ , where  $n_i$  is the C2SS estimate for the  $i^{\text{th}}$  race group,  $n_{i,c}$  is the Census estimate for the  $i^{\text{th}}$  race group, and  $n_{i,c}$  is the grand

total. An overall NDR for race can be estimated by summing the absolute values of those differences over all of the  $j$  race groups:  $\sum (\text{abs}(n_j - n_{j,c})) / n_{j,c}$ .

The second statistic is the Gross Difference Rate (GDR). The GDR is defined as the expected percentage of people whose race group based on the C2SS is different than their race group from the Census. The GDR is estimated by the following formula, summing over the  $j$  race groups:  $1 - \sum (n_{ij}) / n_{i,c}$ . There is no GDR at the individual race group level. Instead, an "Inconsistency Rate" (IR) statistic is calculated, defined for the  $i^{\text{th}}$  race group as  $1 - (n_{i,m} / n_{i,c})$  – the percentage of Census people in the  $i^{\text{th}}$  race group who had a reported race in the C2SS in another race group.

Often, statistics like the NDR and GDR are calculated when one survey is the original measure and the second is an identical reinterview. In that case, if the two interviews are identical, the NDR should be close to zero and the NDR is considered to be a measure of bias (Bureau of the Census, 1993). The GDR/2 is in that case, if any errors are uncorrelated, an estimate of the simple response variance (SRV), the average variability of responses to the same question over repeated trials.

However, there is no claim that the Census and the C2SS were identical. This paper has already documented differences in mail questionnaires, length of form, interviewers, instrument, among other things. There is also no claim that either the Census or the C2SS produce a gold standard from which any deviation is considered an error. Therefore, what do these statistics reveal?

At the individual race group level, the NDR becomes the difference in bias between the two measures (Bureau of the Census, 1993). A positive NDR for a particular race group says that race group is more prevalent in the C2SS than in the Census. Another telling statistic is the ratio of the absolute value of the NDR for the race group divided by the proportion of the population in that race group-- for race group  $i$ ,  $\text{NDR}_i / (n_i / n_{i,c})$ . A ratio of less than one percent indicates relatively low error, a ratio of one to five percent indicates moderate error, and a ratio over five percent indicates a race group with high error (Census Bureau, 1993). The NDR at the overall race variable level also measures the bias differences between C2SS and Census, without saying one is right or better.

Without the assumption of something approaching an exact reinterview, the GDR/2 is not a good estimate of the SRV—it underestimates the SRV. However, it does become a measure of the inconsistency in the reporting of race, as is the IR at the race group level.

What do the NDR and GDR together say? If the NDR is low and the GDR is high for race, then there is a lot of variability in the race responses, but it tends to even out—not biased. In that case, the underestimate of the GDR might not be that great. However, if the NDR

becomes high, especially in terms of the GDR, then there is quite a bit of bias and the GDR does not give much information about the SRV.

Note that since the purpose of this analysis is to measure differences in actual race responses, the NDR and GDR are calculated only using people with a race response in both Census 2000 and the C2SS. If a person did not have a reported race in either the Census or C2SS, that person is not used in the calculation of the statistics.

**RESULTS**

**General**

The most basic result is the cross-tabulation of the Census 2000 race versus the C2SS race category, which is shown in Table 4. The total population of 217,913,003 represents the estimate of matched people for the whole country.

Since the C2SS weights are designed to produce annual totals, and this study includes data from three of the 12 months, the weights were multiplied by four when compiling this table. The total is lower than the total United States population because group quarters were not included (since they were not part of the C2SS), only C2SS respondents are represented, and only matched people were included.

Table 4: Race, Census 2000 versus C2SS, for Matched Persons

C2SS Race	Census 2000 Race								Total
	White	Black	Asian	NHOPI	AIAN	Other	* 2+ Race	Missing	
White	160,026,473	205,734	142,979	17,699	238,091	4,136,622	1,224,962	3,136,279	<b>169,128,839</b>
Black	213,835	21,914,769	24,217	2,331	47,539	134,919	305,385	310,884	<b>22,953,878</b>
Asian	94,708	14,980	6,960,287	3,189	8,397	57,354	196,577	154,064	<b>7,489,557</b>
NHOPI	17,698	744	13,547	172,411	2,232	2,967	41,684	12,972	<b>264,256</b>
AIAN	147,111	21,200	6,481	423	1,033,548	66,565	125,910	38,314	<b>1,439,553</b>
Other	1,590,779	106,300	51,930	4,464	91,350	3,227,323	252,151	1,215,086	<b>6,539,384</b>
2+ Race*	1,231,379	446,879	212,820	24,424	179,648	301,263	1,668,780	180,854	<b>4,246,047</b>
Missing	3,232,177	436,235	131,584	8,085	45,577	883,059	136,580	978,193	<b>5,851,490</b>
<b>Total</b>	<b>166,554,159</b>	<b>23,146,843</b>	<b>7,543,845</b>	<b>233,026</b>	<b>1,646,382</b>	<b>8,810,072</b>	<b>3,952,029</b>	<b>6,026,647</b>	<b>217,913,003</b>

\* The White, Black, Asian, NHOPI, AIAN, and Other rows and columns refer to people reporting that race only. People reporting that race along with other races are in the "2+ Races" row and column.

What does this table say at first glance? Note that the Census estimates that respondents reported 8,810,072 million people with Some Other Race, compared to only 6,539,384 in the C2SS. To contrast, the C2SS had more people reported to be White than Census 2000 did. Note that this trend—more White in the C2SS and more Some Other Race in the Census—is the same one seen in Table 2 in the comparison of the published race distributions between Census 2000 and C2SS.

The advantage of the data in this study is the ability to compare the paired responses for people as opposed to just looking at totals. A large part of the aforementioned difference was due to the fact that there were 4,136,622 people that were reported to be Some Other Race in Census 2000 but White in C2SS, compared to 1,590,779 the opposite way—White in Census 2000 and Some Other Race in C2SS.

Table 5 summarizes the data presented in Table 4. It provides the race distribution for Census and C2SS, the NDR at the race group and total levels, the IR for race groups, and the GDR for race as a whole.

Table 5: Race Distribution and Analytical Statistics, Census 2000 vs. C2SS, All Matched People

Race	Cen %	C2SS %	NDR%	IR, GDR%
White	76.43	77.61	1.29	3.59
Black	10.62	10.53	-0.03	3.22
Asian	3.46	3.44	-0.04	5.11
NHOPI	0.11	0.12	0.01	31.39
AIAN	0.76	0.66	-0.10	26.24
Other	4.04	3.00	-1.26	39.38
2+ races	1.81	1.95	0.12	58.95
Missing*	2.77	2.69	-----	-----
<b>Total</b>	<b>100.00</b>	<b>100.00</b>	<b>2.85</b>	<b>5.80</b>

\* Recall that people missing race in either the Census or the C2SS are not included in the NDR, IR, and GDR statistics.

The NDR for race is 2.85 percent and the GDR is 5.80 percent. Given the difficulty of collecting race data and the subjectivity involved (the respondent determines his or her own race), those figures are reasonable.

However, there are large figures for certain race

groups. The largest NDRs are for White and Some Other Race, which almost balance each other out, and was expected given the previous discussion on Table 4. The GDRs are large for two relatively big groups—Some Other Race and two or more races. Table 4 shows how inconsistent the data were for those two groups. While the Some Other Race group showed a bias in one direction, though, the two or more races responses were inconsistent but relatively balanced.

The ratio of the NDR to the Census percentage is in the low or moderate range (less than five percent) for White, Black, and Asian, and high for the other race groups, with Some Other Race having the highest ratio at 31.10 percent (1.26% / 4.04%).

**Differences by Hispanic Origin**

Given the results of Table 3, which looked at the race distribution for Hispanics, it seems prudent to look at the crosstabulation for Hispanics only. For the purposes of this study, a person was considered to be Hispanic if he or she was reported to be Hispanic (as opposed to edited or imputed) in either the Census or the C2SS.

Table 6 is similar in layout to Table 5 except that only Hispanics are included—an estimated 25.7 million people, 11.79 percent of the matched people. For space reasons, only the White and Other race lines are shown.

*Table 6: Race Distribution and Analytical Statistics, Census 2000 vs. C2SS, Hispanic Matched People*

Race	Cen %	C2SS %	NDR%	IR,GDR%
White	43.08	58.14	14.37	36.74
Other	33.19	24.40	-12.74	38.09
Total	100.00	100.00	28.82	39.33

Table 6 shows that the inconsistency in White versus Some Other Race is especially prevalent for people of Hispanic origin, mirroring the results from Table 3 for the overall race distribution.

Clearly, there were differences in the reporting of race for Hispanics. Are some of those differences related to the data collection mode? As mentioned before, Census used two modes: mail and personal visit (with a paper form), while C2SS used three modes, mail, CATI, and personal visit (using CAPI).

That lead to a comparison of reported race for Hispanics that were collected with the same type of interview in the Census and the C2SS—mail/mail and personal visit/personal visit (PV/PV).

The notable difference is among Hispanic people whose data were collected by an interviewer in both the Census and C2SS. Table 7 is similar to Table 6, but is limited to Hispanics who responded to both Census 2000

and the C2SS via a personal visit interview, an estimated 4.9 million people, 2.25 percent of all people and 19.12 percent of Hispanics.

*Table 7: Race Distribution and Analytical Statistics, Census 2000 vs. C2SS, Hispanic Matched People, Data Collected Via Personal Visit in Both*

Race	Cen %	C2SS %	NDR%	IR,GDR%
White	44.36	68.18	24.57	48.06
Other	44.31	22.82	-23.91	44.91
Total	100.00	100.00	50.48	48.80

Table 7 shows that in Census 2000, the proportions of White and Some Other Race were about the same, while in the C2SS, the ratio was about three to one. The IRs for the race groups and the GDRs are all extremely high, too.

Why might there be an interviewer effect? Recall the Census 2000 interviewers were temporary employees, many new to interviewing, while the C2SS interviewers are permanent Census Bureau employees. Leslie, Raglin, and Schwede (2002) found in a debriefing study of C2SS interviewers that interviewers that did not work on Census 2000 were more likely to probe for another answer if an Hispanic person gave their race as “Hispanic” rather than one of the race groups we were looking for. In addition, some of the surveys the permanent interviews work on do not have a “Some Other Race” option, so they are not as used to having that option.

**Differences by Data Collection Mode**

Table 8 shows the NDRs and GDRs for the matched people, breaking out the results for the mail/mail and PV/PV pairs, as well as for Hispanic/Not Hispanic.

*Table 8: NDRs and GDRs for Race, By Hispanic Origin and Census/C2SS Data Collection Modes*

(Data are in percent)		Data Collection Modes		
Hispanic Origin		Total	Mail/Mail	PV/PV
Total	NDR	2.85	0.35	10.11
	GDR	5.80	2.21	13.51
	% Popul*	100.00	51.99	10.95
Hispanic	NDR	28.82	4.52	50.48
	GDR	39.33	19.74	48.80
	% Popul*	11.79	3.65	2.25
NonHisp	NDR	0.55	0.15	2.01
	GDR	2.16	1.22	4.61
	% Popul*	88.21	48.34	8.70

\* % Popul is the % of study people that fall into that cell

Note that the reporting of race for the mail/mail

people is much more consistent than for the PV/PV people, regardless of the Hispanic origin. Why was that?

First, households were not assigned to have their data collected by personal visit randomly. Their data were collected in the field because in most cases they had not responded to the mail questionnaire. Therefore, these people are the hardest to collect data from. In the Census, as a last resort, interviewers were allowed to collect data from proxies outside the household.

Second, recall the differences mentioned earlier in this paper regarding the interviewers and instruments. The Census used, out of necessity, inexperienced interviewers, and the data were collected via paper and pencil. To contrast, C2SS data were collected in the field by permanent interviewers using a computer instrument.

Those might be reasons there were inconsistent data from people of Hispanic origin—the Census interviewers were more likely to leave the race as “Some Other Race” while the C2SS interviewers were more likely to get an answer that was one of the races we had listed.

**Non-Hispanic Mail/Mail Response Group**

An interesting finding from Table 8, though, is the consistency of Non-Hispanics in the mail/mail response group. This is a large population—almost half of the people in the study, and the NDR and GDR for them is very low. Table 9 gives additional information about the people in that cell:

*Table 9: Race Distribution and Analytical Statistics, Census 2000 vs. C2SS, Non-Hispanic Matched People, Data Collected Via Mail in Both*

Race	Cen %	C2SS %	NDR%	IR/GDR%
White	86.47	85.46	0.01	0.45
Black	7.15	6.95	-0.03	1.34
Asian	3.67	3.62	-0.03	2.36
NHOPI	0.07	0.07	-0.00	16.54
AIAN	0.42	0.43	0.02	21.12
Other	0.10	0.09	-0.00	55.75
2+ races	1.16	1.19	0.05	40.50
Missing	0.96	2.17	-----	-----
Total	100.00	100.00	0.15	1.22

The NDRs are low in this table for almost every race group. The ratio of NDR to Census percentage is in the low range for White, Black, and Asian, and in the moderate level for AIAN, Some Other Race, and 2+ Races. They are barely in the high range for NHOPI, the smallest race group. The high IRs for Some Other Race and 2+ races indicate some inconsistency, but the NDRs indicate that there is minimal bias between the Census and

the C2SS in the reporting of those items.

This is potentially an interesting finding. There is often concern about the consistency of race reporting, but these data indicate that for a large share of the population—Non-Hispanics who are willing to fill out the mail forms—race reporting is consistent with the exception of people reporting two or more races.

**CONCLUSIONS**

Race is considered to be a very difficult data item to collect consistently. By the government’s definition, a person’s race is determined by the person—there are no biological definitions. However, this paper has identified some factors that seem to be related to the consistency of race reporting.

The Census Bureau seems to have a problem collecting consistent race data for people of Hispanic origin, especially people whose data are collected via personal visit. Part of that may be due to interviewer effects, and part of that may be due to proxy reporting from outside of the household. Only a study that randomly assigns people to a certain data collection mode can tease out those effects.

However, this paper found a large subpopulation for which race seems to be reported consistently—Non-Hispanics who respond to mail surveys, with the important exception of two or more races. This would suggest that future research into race reporting might be able to focus on certain subpopulations of the country.

**REFERENCES**

Bates, N., Martin, E., DeMaio, T. and de la Puente, M., 1995. “Questionnaire Effects on Measurements of Race and Spanish Origin,” in *Journal of Official Statistics*, Volume 11, No. 4, pp. 433-459.

Bureau of the Census, 1983. “1990 Census of Population and Housing, Evaluation and Research Reports, Content Reinterview Survey”, Report 1990 CPH-E-1, U.S. Government Printing Office, Washington DC, October 1993.

Leslie, T., Raglin, D., and Schwede, L., 2002. “Understanding the Effects of Interviewer Behavior on the Collection of Race Data”. Presented at the American Association for Public Opinion Research Conference, St. Petersburg, FL, May 2002.

Martin, E., DeMaio, T. and Campanelli, P., 1990. “Context Effects for Census Measures of Race and Hispanic Origin,” in *Public Opinion Quarterly* 54: 551-566.