Comparison of Edit and Imputation[1] Procedures for the Question on Hispanic Origin: 1990 Census and Census 2000[2]

Dr. Arthur Cresce, U.S. Census Bureau[3]

Acknowledgments

I would like to acknowledge the assistance of Roberto Ramirez who produced most of the statistical tables that were the basis for the empirical analysis in this paper. He spent a great deal of time producing these tables from the detailed 1990, Census 2000, and Census 2000 Supplemental Survey (C2SS) files and these tables were essential to making key points in this paper. I also would like to acknowledge Dr. Gregory Spencer who provided valuable assistance in the development of this paper and Dr. Aref Dajani who reviewed this paper and provided extremely useful suggestions to improve it.

Executive Summary

Comparison of the 100-percent edit and imputation procedures for the 1990 Census and Census 2000 reveals differences between the two procedures. (See Figure 1.) In general, the 1990 100-percent procedures were not as rigorous as the corresponding Census 2000 procedures in assigning an origin. One significant difference between the specifications for the two procedures is the use of surname-assisted hot decks in Census 2000.

An extremely important context for understanding the impact of these differences is the fact that the number of imputations for the origin question dropped by 34 percent between 1990 and 2000. This translated into a drop from 25.5 million imputations in 1990 to 16.8 million imputations in 2000. In addition to the drop in overall imputations, there was a fundamental shift in the type of imputation made. In 1990, 75.6 percent of imputations occurred through the "hot deck" method that relied on the reporting of people who lived nearby. By contrast, only 41.2 percent of imputations required hot deck imputation in Census 2000. This is an important point, because of the techniques used in the 1990 census (imputation based on other information provided by the respondent, imputation from other household members, and hot deck imputation), hot deck imputation was the least reliable. We can attribute this improvement, in large part, to moving the question on origin before the question on race.

There is strong evidence that the less restrictive 1990 100-percent edit and imputation procedures and greater reliance on hot deck imputation, combined with a much higher level of nonresponse to the Hispanic origin question in 1990, may have resulted in misclassifying at least 161,000 people as Hispanic. We did not attempt to run the Census 2000 edit and imputation program on 1990 data because the Census 2000 program used surname-assisted hot decks and there was no capture of surnames on short forms in the 1990 census. However, we believe the Census 2000 100-percent procedures would have misclassified fewer people as Hispanic than did the 1990 program. This evidence of misclassification, however, should not be construed to imply that there was an overcount of Hispanics in the 1990 census. It merely indicates that among the enumerated population, our 1990 100-percent edit and imputation program incorrectly edited as Hispanic some people who probably were not Hispanic.

Philosophy of Edit and Imputation Procedures

In any imputation method, imputed values may differ (sometimes significantly) from what would have been obtained had the information been reported by the respondent. Edit and imputation techniques are designed to provide the best possible estimate of the probable response given the best information available. For example, if the respondent did not provide an origin, the procedure first checked to determine if the person indicated that he or she was Hispanic in the question on race (close to half of Hispanics provided an Hispanic ethnicity in the question on race). If an origin could not be obtained from race, then the procedures attempted to impute an origin from other people in the household (according to a hierarchy of household relationship) under the assumption that people living in the same household would tend to have the same origin. If an origin could not be obtained from within the household, an origin was assigned by hot deck imputation under the assumption that people of the same origin tend to live in close proximity to each other. To the extent that these assumptions did not hold for a given person or household, imputed values might have differed from what would have been obtained had the information been obtained directly from the respondent.

Edit and imputation procedures attempt to rely as much as possible on sources of information about which there is the most confidence (other information provided by the respondent or responses of other household members) and to rely less on procedures such as hot deck imputation. However, hot decks, depending on the

---

[1] In this report, "edit" refers to revising or imputing a response based on information provided directly by the respondent. "Imputation" refers to assigning a response based on the response of other people in the same household or the response of people in neighboring households.

[2] 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.

[3] Mail address: 4700 Suitland Road, Room 2024, Suitland, MD 20233
Email: arthur.r.cresce.jr@census.gov

sophistication of matching criteria for donors and recipients, can improve the accuracy of imputation by matching donors and recipients according to one or more key characteristics. For example, in the 1990 census, origin hot decks used race as a matching variable for donors and recipients. In contrast, Census 2000 used not only race, but also age and whether the surname was Spanish or not Spanish, as matching variables. We believe these additional variables improved the accuracy of origin imputation from the hot deck.

Comparison of Edit and Imputation Procedures for Hispanic Origin

Summary of Differences

Figure 1 summarizes the key differences between the edit and imputation procedures for the Hispanic origin question in 1990 and 2000. First, while multiple responses were not allowed in either census, Census 2000 allowed for the data capture of more than one response and the edit and imputation procedures assigned one origin. In the case of multiple non-Hispanic or multiple Hispanic responses, a respondent remained non-Hispanic or Hispanic, respectively. However, in the case of a conflicting Hispanic/non-Hispanic response, an attempt was made to resolve this conflict by using other information provided by the respondent (for example, an Hispanic response in the race question), responses of other people in the household or people living nearby, who are of the same race.

Census 2000 edit and imputation procedures also differed from the 1990 procedures in how origin could be assigned from other people in the household. In 1990, anyone in the household could donate an origin regardless of their race. By contrast, Census 2000 rules only allowed other household members to donate an origin if the person needing an origin and the donor had the same race.

One of the most important differences between the two procedures was how "hot deck" imputation was implemented.[4] In 1990, hot deck values were stored and assigned by the race of the donor and recipient. In Census 2000, hot decks additionally were controlled by four broad age groups.

More importantly, Census 2000 origin hot decks were further differentiated by whether the donor (and recipient) had a Spanish or non-Spanish surname. Use of surname in storing and assigning an origin was one of the most important innovations implemented in Census 2000 in that it allowed a much more precise method for assigning an origin from a hot deck. This innovation was cited in a recent evaluation of having a "profound" impact on the assignment of origin.[5]

Finally, if both race and Hispanic origin were not reported, the edit attempted to assign both a race and an origin from another donor (both within-household imputation and hot deck imputation). The 1990 procedures assigned race and origin independently of each other, thus increasing the possibility of creating race/origin combinations that were not common in the population.

Context for Comparing Edit and Imputation Procedures

Before assessing the impact of these differences on the Hispanic origin population, it is important to understand the differing contexts within which each edit operated. One of the hallmarks of the Hispanic origin question in 1990 was the relatively high level of nonresponse. Imputation rates[6] for the 1990 census were almost twice as high in Census 2000 (10.4 percent versus 5.6 percent). What is striking is that the range of imputation rates by region narrowed considerably from 1990 to 2000. In 1990, the rates ranged from 7.2 percent in the West to 11.8 percent in the Northeast – a difference of 4.6 percentage points. Among states and the District of Columbia, the range was even wider with Idaho having the lowest percent (4.2 percent) and the District of Columbia having the highest (18.3 percent) – a difference of 14.1 percentage points. In Census 2000, by contrast, the range by region was much narrower, with the Midwest having the lowest rate (4.7 percent) and the South having the highest rate (6.0 percent) – a difference of only 1.3 percentage points. By state, Nebraska had the lowest rate in Census 2000 (3.5 percent), while the District of Columbia had the highest rate (11.0 percent) – a difference of 7.5 percentage points. It is clear that the biggest improvement in these rates occurred for states that had high imputation rates in 1990. This dramatic improvement in response can be attributed in large part to the placement of the Hispanic question before the question on race in Census 2000.

At the national level, hot deck imputation was the largest source of origin response after "reported origin." This means that for a substantial proportion of the population (8.0 percent), no one in the household

[4] "Hot deck" imputation involves the assignment of values from a set of stored values that are constantly updated as each person's data record is processed. A hot deck is usually the last procedure used when a value cannot be assigned either from information provided by the person or from other people in the household. In the case of race and origin, hot deck imputation is used most often when no one in the household has provided a response to a particular question.

[5] Summary provided by Yves Thibaudeau, Statistical Research Division, March 31, 1999 concerning evaluation of editing of origin in the 1998 Census Dress Rehearsal.

[6] Imputation rates represent the rate at which responses were imputed based on responses of others within the household or from people living nearby (also called "hot deck" imputation).

answered the Hispanic origin question. This relationship held for all states.

For the 1990 Hispanic population, there was about equal reliance on "within-household" and "hot deck" imputation, with some regions and states having a higher proportion of within-household imputation. This is not surprising since the question is primarily oriented to the Hispanic population. By contrast, the proportion of responses coming from hot deck imputation for non-Hispanics was much higher than that from within-household imputation.

One of the most important changes made to the Hispanic origin question in Census 2000 to address the problem of nonresponse was to shift the order of the Hispanic origin and race questions. In the 1990 census, the race question appeared first and the Hispanic origin question appeared several questions later. It seems clear that after answering the question on race, many people felt that the Hispanic origin question did not apply and simply skipped the question. Shifting the order of the questions in tests conducted before Census 2000 seemed to improve overall response to the Hispanic origin question with some increased nonresponse to the question on race.

It is very clear from the data that not only the level of nonresponse was reduced but also that the relative contribution of within-household and hot deck imputation was much more balanced for non-Hispanics in Census 2000 than in the 1990 census. More importantly, imputation from surname-assisted hot decks overall was greater than imputation from non-surname-assisted hot decks. For example, among non-Hispanics, imputation from surname-assisted hot decks was about three times the level of imputation from non-surname-assisted hot decks (2.0 percent compared to 0.6 percent).

The impact of surname-assisted programs is clearly more dramatic when observing the source of imputations. Overall, surname-assisted hot decks represented 31.4 percent of all imputations, while non-surname-assisted hot decks accounted for only 9.6 percent of all imputations. For Hispanic imputations, surname-assisted hot decks overall represented 8.1 percent of all imputations while non-surname-assisted hot decks represented about 4.0 percent. For non-Hispanics, surname-assisted hot decks provided 36.9 percent of all imputations, while non-surname-assisted hot decks provided only 10.9 percent of all imputations. In West Virginia, Kentucky, Mississippi, Arkansas, and Tennessee - all States where the there are few Hispanics - the ratio of surname-assisted hot deck imputations to non-surname-assisted imputations is at least 4 to 1.

It is clear that there was a significant increase in Census 2000 in the level of substitution, from 0.7 percent of the population in households in 1990 to 1.2 percent of the total population in Census 2000. Substitution occurs when there are no data for anyone in the housing unit, and we use data from a neighboring household of similar size, using the hot deck method, to impute characteristics for the people in that housing unit. Given that the same basic method was used in both censuses, there is no reason to believe that the procedure itself created any upward or downward bias in assigning origin in 1990 and 2000 or there could be bias in both censuses.

The percent substituted was slightly higher for the Hispanic population (1.6 percent) than for the non-Hispanic population (1.2 percent) in Census 2000. There was a similar pattern in 1990, however, but at a lower level - the percent substituted for the Hispanic population (0.9 percent), slightly higher than that for the non-Hispanic population (0.6 percent). In addition, it is also clear that substitution played a much larger role in the source of imputation of origin in 2000, with substitution constituting about 20 percent of imputations overall. Interestingly, the share of substitution was higher for the non-Hispanic population (21.1 percent) than for the Hispanic population (17.5 percent). By contrast, in 1990 the share of substitution in total imputations was much higher for Hispanics (11.0 percent) than for non-Hispanics (5.9 percent). The reasons for the increase in substitution will be part of the Census Bureau's evaluation of Census 2000.

Finally, to put these results in a broader perspective, the results from the Census 2000 Supplemental Survey (C2SS) show that the trend toward improved response to the origin question is continuing. Editing procedures were basically the same for Census 2000 and the C2SS, except that there was no substitution in the C2SS. In particular, imputation rates are lower for the total population and for the Hispanic and non-Hispanic populations in the C2SS than in Census 2000 and in 1990. There was an even greater reliance on surname-assisted hot decks in the C2SS than in Census 2000, with the C2SS showing a much greater reliance on surname-assisted hot decks for the non-Hispanic population than for the Hispanic population. It should be noted, however, that the level of response in C2SS was improved through the use of field follow-up procedures for people who did not fully answer the questions on the questionnaire, a procedure that was not used in Census 2000.

Impact of Editing on Hispanic Origin Population in 1990

In the 1990 census, there was an unusually high level of dependence on hot deck imputation because many of the people needing an imputed origin had no reported origin for anyone in the household. This greater reliance on hot deck imputation, combined with a relatively high level of nonresponse, meant that most imputations came from the hot deck, especially for the non-Hispanic population. For example, in 1990 78.9 percent of non-Hispanic imputations came from a hot deck, excluding substitutions. By contrast, only 29.9 percent of Hispanic imputations came from a hot deck, again excluding substitutions.

Concerns about the impact of 1990 edit and imputation procedures emerged when the results of the sample data processing, including a separate edit and imputation for sample questionnaires, became available. The Hispanic origin question on the sample form was edited in sample processing independent of the 100-percent edit and imputation program. Although the basic structure of the two procedures were the same, the edit and imputation procedures for the Hispanic origin question during sample processing differed in a very important way from those used in 100-percent processing. Unlike the 100-percent procedures, sample procedures made use of the rich source of ethnic-related questions from the sample form (ancestry, place of birth, language spoken at home) that could assist in imputing for nonresponse. The use of ethnic-related information, combined with a higher response rate for the Hispanic origin question on the long form, meant a much lower dependence on hot deck imputation.

The estimate of the Hispanic origin population that resulted from sample processing was about 454,000 below the total of Hispanics obtained from 100-percent processing with the 100-percent total exceeding the sample estimate for most states. This difference existed despite the fact that sample estimates were controlled to 100-percent totals, including race and Hispanic origin.[7]

Thompson (1991) addressed this difference and the difference between 100-percent totals and sample estimates for the American Indian population. Thompson attributed the difference between 100-percent totals and sample estimates for Hispanics primarily to 1) undersampling of Hispanics, 2) a form of imputation bias, and 3) different data processing procedures.[8] His analysis, however, did not quantify how much each factor contributed to this difference.

The "imputation bias" to which Thompson's analysis referred is directly related to the focus of this analysis. Thompson noted that the nonresponse for the Hispanic question on the short form was 10 percent while the nonresponse rate for the same question on the sample form was only 4 percent. This difference was due partly to the fact that during data collection all sample forms were subject to content edit follow-up (field follow-up of cases where the number of non-reported items exceeded a certain threshold). By contrast, only 10 percent of short forms were subject to content edit follow-up.

Thompson reasoned that Hispanics were more likely to answer the Hispanic origin question than were non-Hispanics, making the donor pool more heavily Hispanic than it would have been had both Hispanics and non-Hispanics reported. If the nonresponse rate for the Hispanic question was high, there was an increased risk that an Hispanic origin would be disproportionately assigned. Evidence of this comes from Del Pinal (1994) who noted that the 1990 edit and imputation procedures tended to increase the overlap between various racial groups and the Hispanic population. For example, although there were very few Black Mexican origin persons, about 62 percent of Black Mexicans were created by the edit and imputation procedures.[9] Not surprisingly, the Black population had a much higher nonresponse rate (18.4 percent) in the Hispanic origin question than did the White population (9.6 percent). The corresponding nonresponse rates for American Indians and Alaska Natives and Asians and Pacific Islanders were 10.2 percent and 9.7 percent, respectively. All these rates were still much higher than the nonresponse rates for other 100-percent questions such as race, age, gender and household relationship – all of which had nonresponse rates below 3 percent – and increased the possibility of a misclassification of respondents as Hispanic. To give a sense of the potential impact on the data, a net misclassification of only 0.1 percent of nonresponses as Hispanic out of a total of 24 million needing an origin would result in a net increase of 240,000 Hispanics.

To attempt to quantify at some minimal level the impact of the potential misclassification of responses as Hispanic, we obtained records from the sample edited detailed file (SEDF) for 1990. On these records, we had not only the origin value from sample processing (along with its imputation flag to indicate whether the value was reported or imputed) but also the origin value from 100-percent processing along with its corresponding imputation flag. In particular, we were interested in

---

[7] Although efforts are made to control the weighting by race and Hispanic origin in each weighting area, there is no guarantee that these weighting control totals can be maintained in each area because each control total in the weighting matrix had to meet a certain minimum threshold.

[8] In 1990 processing for the Hispanic origin question, only optical marks, but no write-in responses, were captured. Thus, people providing a write-in response without filling the "Other Hispanic" circle were treated as a nonresponse in the 100-percent edit and they could have been assigned either as Hispanic or not Hispanic. People providing a write-in response and marking the "Other Spanish/Hispanic" circle would have been "Other Spanish/Hispanic" in the 100-percent edit, but would have been either Hispanic or not Hispanic in the sample edit depending on whether the write-in response was Hispanic or not Hispanic in sample coding operations.

[9] The percentages and rates in this paragraph were derived from special 1990 files containing only household records and excludes records from the group quarters population (such as college dorms, prisons, military bases, and nursing homes).

determining how people who received an imputed origin in the 100-percent edit had their origin imputed in the sample edit. For the purposes of this analysis, the results of the sample edit are considered the standard for accuracy because sample editing procedures made use of data from additional ethnic-related questions (ancestry, place of birth, and language spoken at home) not available on the short form.

This analysis showed that the 100-percent edit produced a net of about 181,000 more Hispanics than did the sample edit when origin was imputed both in 100-percent and sample editing procedures. This net difference in edit outcomes represented only 1.2 percent of the 8.6 million people for whom origin was allocated in both 100-percent and sample processing.

If we take into consideration also the situations in which we imputed a value in the 100-percent procedures but did not impute a value in the sample procedures, the 100-percent edit produced a net overall of about 161,000 more Hispanics than did the sample procedures.[10] Assuming that the sample edit and imputation process is more accurate, the 100-percent edit appears to have imputed as Hispanic a net total of 161,000 people who were probably not Hispanic. However, this number represents only 1.8 percent of all people whose origin was imputed. It is also important to keep in mind that both edit procedures agreed on the edit outcome 96 percent of the time.

It is clear from this analysis that the impact of this potential misclassification is different by race. The apparent degree of misclassification of Hispanics (measured by taking the ratio of "Hispanic – 100%; Not Hispanic – Sample" to "Not Hispanic – 100%; Hispanic – Sample") appeared to be much greater for Blacks (10.0) and Asian and Pacific Islanders (13.1) than for Whites (4.4). Analysis of the unweighted data shows the same pattern, but slightly lower ratios for each group. This finding is consistent with Del Pinal's finding that certain race/Hispanic combinations were more significantly affected by the editing procedures.

It is important to keep in mind that the estimate of 161,000 is probably a lower bound because these data were obtained from sample forms that had a lower nonresponse rate and had much more ethnic-related information than did short form questionnaires. It is possible that the level of misclassification would be higher among the population that received only the short form, which experienced a higher nonresponse rate for origin than did the sample form. However, it is unlikely that the upper bound would be as high as the total difference between the 100-percent and sample totals (454,000) because: 1) sample processing changed about 262,000 responses from "Other Spanish/Hispanic" to not Hispanic[11] and 2) to an unknown degree there was undersampling of Hispanics for which the sample weighting procedures did not compensate.

It is also very important to keep in mind that the impact on the overall total Hispanic population was very small. Overall, this net difference (161,000) represented only 0.7 percent of the total Hispanic population.

Finally, this evidence of misclassification should not be construed to imply that there was an overcount of Hispanics in the 1990 census. It merely indicates that among the enumerated population, our 100-percent edit and imputation program incorrectly edited as Hispanic people who probably were not Hispanic.

Impact of Edit and Imputation Procedures on Hispanic Origin Population in Census 2000 There are no comparable data available at this time from Census 2000 to perform the same type of analysis that was conducted on the 1990 census edit and imputation procedures. However, it is very clear that the Census 2000 procedures operated in an environment that was profoundly different from that in which the 1990 procedures operated. We believe that significantly reduced nonresponse to the question, combined with more restrictions on the conditions under which origin could be assigned to an individual, probably has led to a much lower level of erroneous imputations as Hispanic (or non-Hispanic).[12] At the same time, innovations, such as the surname-assisted hot deck, has improved the accuracy and, therefore, the quality of data from the Hispanic origin question.

Conclusion

There were substantial differences in the edit and imputation procedures between the 1990 census and Census 2000 for the Hispanic origin question. The most

[10] This was possible because we only captured optical marks in the 100-percent data processing and a person could have written in a response without marking any circles. Although the write-in entry could have been either an Hispanic or a non-Hispanic entry, most of the time the entry was Hispanic.

[11] Based on the fact that the respondent provided a non-Hispanic response in the write-in space.

[12] Another example of this is how we handled situations in which a respondent indicated that he or she was Hispanic and non-Hispanic. This situation occurred about 700,000 times nationally. Instead of assuming that all such people should be Hispanic, we looked at information provided by the respondent (such as the reporting of an Hispanic origin in race), information provided by others in the household or the hot deck, to adjudicate these situations. We discovered that about half of the people were assigned as Hispanic and half were assigned as not Hispanic.

important of these was the use of surname-assisted hot decks in Census 2000. These hot decks allowed for much greater precision in assigning an origin from neighboring housing units when no one in the household answered the question. Furthermore, there was a dramatic improvement in response to the Hispanic question in Census 2000, thus reducing the need (relative to 1990) for providing a response through edit and imputation procedures. In fact, there is evidence from 1990 that the combination of higher nonresponse, greater use of hot deck procedures, and lack of the benefit of surname-assisted hot deck procedures (surname capture was not done in 1990 for all census forms) led to some misclassification of people as Hispanic.

We will continue our analysis of the quality of Census 2000 origin data as sample data and data from other evaluation studies become available.

For a more extensive analysis, including detailed tables, contact the author at the address noted in footnote 3 on page 1.

Bibliography
Sources cited in report:

Del Pinal, Jorge. "Social Science Principles: Forming Race-Ethnic Categories for Policy Analysis." Paper presented at the "Workshop on Race and Ethnicity Classification: An Assessment of the Federal Standard for Race and Ethnicity Classification," National Research Council, Commission on Behavioral and Social Sciences and Education, Committee on National Statistics, February 18, 1994.

Thompson, John H. "Difference Between Complete Count Figures and Sample Estimates by Race/Origin." Memorandum from John H. Thompson, Chief, Statistical Support Division to Charles D. Jones, Associate Director, Decennial Census, December 19, 1991.

Figure 1. Differences Between Census 2000 and 1990 Census Edit and Imputation Procedures

Reporting of more than one origin

Census 2000 - All responses were retained for research purposes. Resolution to one origin was accomplished using a set of edit rules.

1990 Census - Multiple responses were not retained. Data capture and data processing retained only one origin.

Within-household imputation

Census 2000 - Assignment of origin was based on another person in household (according to a pre-defined priority order of household relationship) with the same race.

1990 Census - Race match was not required.

Surname-assisted hot decks

Census 2000 - Three separate hot decks were used:
1) surname is Spanish
2) surname is not Spanish
3) surname is not clearly Spanish or not Spanish or surname is not reported.

1990 Census - Separate hot decks were not used.

Joint allocation of race and origin

Census 2000 - If both race and origin were not reported, an attempt was made to assign both race and origin from the same donor within the household. If hot deck assignment was required, both race and origin always were assigned from a single donor.

1990 Census - Each value was assigned independently of the other. Race and origin might not necessarily have come from the same donor.