Population Division

ANALYSIS OF MULTIPLE ORIGIN REPORTING TO THE HISPANIC ORIGIN QUESTION IN CENSUS 2000

by

Roberto R. Ramirez

Working Paper No. 77

U S C E N S U S B U R E A U

Helping You Make Informed Decisions

Acknowledgments

The author wishes to thank Art Cresce, and Jorge del Pinal in Population Division for their overall guidance for this paper. Content review was provided by Frank Hobbs and Campbell Gibson.

Table of Contents

FIGURES:

TABLES:

ANALYSIS OF MULTIPLE ORIGIN REPORTING TO THE HISPANIC ORIGIN QUESTION IN CENSUS 2000

Introduction

For the first time in U.S. Census history, respondents in Census 2000 were given the option of reporting more than one category in the question on race. Thus, for example, individuals could report multiple racial categories, such as White and Black, or Black and Asian, or White, Black and Asian.[1] The implications of allowing people to identify with several races has received much attention by the general public and especially by the research community.

By contrast, far less attention has been paid to the reporting of multiple origins on the Hispanic origin question.[2] Unlike the race question, only one answer was solicited in the Hispanic origin question; there were no specific instructions on the Hispanic origin question allowing individuals to identify with more than one origin although this type of reporting did occur. Some responses included specific detailed Hispanic combinations such as Mexican and Cuban (Multiple Hispanic detailed origins), or Hispanic and Non-Hispanic combinations such as Non-Hispanic and Mexican (Part-Hispanic). In addition, other responses observed in the data include people not providing an origin at all (blanks) and people responding with a single origin term such as "American."

Census 2000 was a unique Census in that unlike other previous Censuses in which only one origin response to the Hispanic origin question was captured and processed, all different types of responses (including multiple check boxes and multiple write-ins) were captured for research and evaluation purposes. In fact, the data processing procedures for Census 2000 were specifically designed to capture all different types of responses. Those additional data collected on the Hispanic origin question have allowed the Census Bureau to evaluate and review the nature of different types of responses to the Hispanic origin question more closely. It is hoped that through this research effort, improvements will be made to the Hispanic origin question.

Very little was known about the different types of responses to the Hispanic origin question until recently. Results from Census 2000 do indicate that there was a wide range of reporting although the vast majority of people identified with one ethnic group. However, there still was a number of people across the country who elected not to identify with one ethnic group

---

[1] The race classifications in the Census are social-political constructs and should not be interpreted as being scientific or anthropological in nature. For more information see, Elizabeth M. Grieco and Rachel C. Cassidy, 2001, Overview of Race and Hispanic Origin:2000, U.S. Census Bureau, Census 2000 Brief, C2KBR/01-1. This report is available on the U.S. Census Bureau's Internet site at www.Census.gov/prod/2001/pubs/c2kbr01-1.pff.

[2] Unlike the race question, there were no specific instructions or provisions for selecting more than one Hispanic origin in the Hispanic question although for Census 2000, this information was captured and processed in the One Hundred Census Unedited File or (HCUF). For those individuals who had selected more than one Hispanic origin, (Mexican and Cuban for example), they were assigned a Hispanic code of '291' and placed under the "Other Hispanic" category in the code list. No such provisions were made in the 1990 Census for those individuals who had reported more than one Hispanic origin. For those respondents who had reported more than one Hispanic origin they were generally coded into the largest Hispanic origin group but some may have been coded into the smallest Hispanic group. For example, if someone had reported that they were Mexican and Cuban in 1990, they were coded in the Mexican group although no hard numbers exits since this information was not captured in 1990.

and responded by providing multiple ethnic responses or none at all. The implications of understanding this kind of reporting is very important; especially to better understand how people of Hispanic and Non-Hispanic background identify ethnically in the United States today. In fact, the Hispanic/Latino population represent one of the fastest growing minority groups nationally as evidenced from the results of Census 2000 in which over 35 million Hispanics were counted (a 58 percent increase since 1990). The vast majority of Hispanics in the United States are of Mexican origin (58.5 percent), followed by Other Hispanic origin (28.4 percent), Puerto Rican (9.6 percent), and Cuban (3.5 percent, Betsy Guzman, 2001).

Purpose

The purpose of this paper is to examine the reporting of multiple origins in the Census 2000 question on Hispanic origin. We begin with a brief discussion of the three classifications of responses to this question (no response, single response, and multiple response). The main focus of this paper will be on multiple origin reporting using unedited and edited data from Census 2000. After a brief discussion of how multiple responses were edited in Census 2000, we will examine the various components of multiple origin reporting. Particular attention will be paid to people who reported combinations of Hispanic and non-Hispanic (Part-Hispanic) and people who reported Multiple Hispanic and multiple non-Hispanic origins. The geographic distribution of multiple origin reporting by type is shown for the total United States as well by Census regions and States. Demographic characteristics of the multiple origin population and the development of a multivariate logistic regression model were used to describe and identify associations among selected demographic variables with multiple origin reporting. Additional background information from respondents who were classified as Part-Hispanic will be further examined using ethnic-related data such as ancestry, place of birth, and language from the Census long form to check for origin reporting consistency or inconsistency. We conclude this working paper with summary of our findings and some recommendations concerning the continued collection on multiple origin reporting.

Background

In 1977, the Office of Management and Budget (OMB) issued the first set of federal government standards for the collection and presentation of race and Hispanic origin data for all governmental agencies. The standard specified four categories for data on race (White, Black, Asian or Pacific Islander, and American Indian or Alaskan Native) and two categories for data on Hispanic ethnicity ("Hispanic origin" and "Not of Hispanic origin").[3] Throughout these standard categories reflected the growing need by federal agencies to have uniform categories on particular population groups used for enforcement of civil rights and voting rights laws

[3] Hispanics may be of any race. See Statistical Policy Directive no. 15, "Race and Ethnic standards for Federal statistics and administrative reporting: May 1977 (see Appendix). " <http://www.whitehouse.gov/omb/fedreg/notice_15.html>.

implemented earlier in the decade. Since then, there have been a number of demographic shifts in the population that have warranted a change in the standards to better reflect the growing diversity in the United States. Thus, in 1997, the federal standards underwent major revisions which allowed multiple race reporting by respondents and provided guidelines in how to present such data to the public.[4] These changes were influenced by the results of four national sample surveys that examined the way people reported race and ethnicity in the general population: The May 1995 Current Population Survey (CPS), Supplement on Race and Ethnicity; the 1996 National Content Survey (NCS); the 1996 Race and Ethnic Targeted Test (RAETT); and finally, the 1997 National Health Interview Survey (NHIS). These four surveys were conducted in the 1990s to test a variety of research questions ranging from testing the effects of separate questions on race and ethnicity with or without a "multiracial" category and the effects of a combined question on race and Hispanic origin (U.S. Bureau of the Census, 1996, 1997). However, none of these surveys tested the possibilities of allowing more than one response to the question on Hispanic origin.

We know that racial/ethnic diversity is increasing in the United States but we know very little about the factors that affect the selection of a single Hispanic/non-Hispanic category versus the selection of several Hispanic and/or non-Hispanic identifiers.

Data and Methods

The questions on Hispanic origin used from 1970 to 2000 were not designed to obtain information on reporting of multiple origins. Thus, very little is known about this population. However, there is increasing interest in this topic. In response to recommendations to include reporting of people who we call "Part-Hispanic", the Office of Management and Budget encouraged the collection of data on multiple origin reporting to research the issue[5]. As a result of the OMB recommendation, the Census Bureau for the first time captured multiple responses to the question on Hispanic origin. During the processing and editing of the unedited census data, single and multiple Hispanic and non-Hispanic responses were captured using a three-digit code[6]. This is further discussed in the section on editing responses from the question on Hispanic origin. All unedited data from the Hispanic origin questions were retained for research purposes.

---

[4] See "Revisions to the standards for the classifications of federal data on race and ethnicity. Federal Register: Oct. 30, 1997"at the following website: <www.whitehouse.gov/omb/fedreg/1997standards.html>.

[5] See "Revisions to the standards for the classifications of federal data on race and ethnicity. Federal Register: Oct. 30, 1997"at the following website: <www.whitehouse.gov/omb/fedreg/1997standards.html>. Under " Topics for further research."

[6] For complete details, see the Hispanic origin code list in the technical documentation for Summary File 1 available at <www.Census.gov/prod/cen2000/doc/sf1.pdf>.

Data regarding the respondents' basic demographic and housing characteristics were collected from the 100-percent short-form questionnaire. Unedited data files were prepared by the Decennial Systems and Contracts Management Office (DSCMO) using data capture specifications for the Census 2000 short form. Two state specific data files were created using the Statistical Application Software Package for each state (the Hundred Percent Census Unedited File, HCUF, and the Hundred Percent Edited File, HDF) for data review and imputation purposes. A SAS program was written to merge the HCUF and the HDF into one national data file (HCUF/HDF merged file). This final file consisted of both unedited and edited data for every person enumerated in the Census. Most of the analysis conducted in this paper is based on this merged national 100 percent data file. The HDF has been edited and imputed for missing values according to guidelines set by the Census Bureau subject matter experts. The HCUF was used to identify the different types of responses to the Hispanic origin question because the original unedited responses were preserved. In order to examine the ethnic background of Part-Hispanics, long form (sample data) variables such as ancestry, place of birth, and language were added to the national merged file by linking the final HCUF/HDF records with records in the Sample Edited Detailed File (SEDF). The United States population (housing unit and group quarters population) is the population universe used in this study. Puerto Rico and U.S. Island Areas are excluded.

The Hispanic origin question

In Census 2000, Hispanic origin was ascertained by the question: **"Is this person Hispanic/Spanish/Latino"**? The instructions of the Hispanic origin question stated: *Mark the "no" box if not Spanish/Hispanic/Latino* (see Figure 1). The Hispanic origin question consisted of five check boxes and one write-in line. The write-in line was provided for those respondents who wanted to write in another type of Hispanic origin (e.g., Dominican, Columbian, Venezuelan...etc...) other than the Hispanic groups already displayed in the check boxes. Up to two write-in entries were captured during Census 2000. The Hispanic question was also placed immediately before the race question. The Hispanic origin question first appeared in 1970 on the 5-percent sample long form and then was asked of everyone in the general population when it was placed on the short form in the 1980, 1990, and 2000 censuses.[7] There have been number of changes to the Hispanic origin question since the last census in 1990. In brief, the changes have centered around:

Format changes: the 1990 short form used a matrix format for 100 percent items, while the 2000 short form used individual person spaces.

Resequencing of race and Hispanic origin questions: In 1990, race preceded Hispanic origin by two questions; in 2000, Hispanic origin preceded race.

[7] The question on "Spanish origin or descent" (as it was known back in 1970) was only asked in the Southwestern States of the United States (California, Arizona, New Mexico, Texas) and in New York and Florida.

Question wording: In 1990, the origin question was, "Is this person of Spanish/Hispanic origin? Fill ONE circle for each person." In 2000, the question was, "Is this person Spanish/Hispanic/Latino? Mark the 'No' box if not Spanish/Hispanic/Latino."

Use of examples: In 1990, examples were printed above the box for "other" write-ins: "Yes, other Spanish/Hispanic (Print one group, for example: Argentinean, Colombian, Dominican, Nicaraguan, Salvadoran, Spaniard, and so on.)" In 2000, the examples were dropped: "Yes, other Spanish/Hispanic/Latino -Print group." (Martin, 2002).

This paper will not address the effects of these changes on the type of responses to the Hispanic origin question. Other research, however, has addressed these issues. For example, there is some evidence indicating that the way the Census 2000 Hispanic origin question was designed (i.e., wording, omission of specific Hispanic origin examples) may have contributed to a significant number of people reporting general Hispanic terms such as "Hispanic" and "Latino" instead of reporting a specific Hispanic origin group such as Colombian or Dominican (Martin, 2002). This same study also indicated that the re-sequencing of the race and Hispanic questions appeared to have lowered nonresponse to the Hispanic origin question. Other recent research conducted on the Hispanic question, however, supports the conclusion that question wording and format changes affected reporting, but also suggests that other factors such as generation differences (e.g., first generation vs. third generation) and term preferences may explain why people chose to identify with a general Hispanic term instead of a specific one (Cresce and Ramirez, 2003). For example, according to the results of the aforementioned study, a total of 5.7 million people choose a general Hispanic term to define their origin in Census 2000. Of these, 3.1 million people (54 percent) were found to have a more specific Hispanic origin group based on their responses to long form questions on place of birth or ancestry. However, nearly half of all the people with a more specific Hispanic origin group, (1.6 million out of 3.1 million for example), were of Mexican origin. This is an interesting finding because this occurred despite a check-box category in the question for Mexican origin. Further research is needed into why individuals of Mexican origin elected to report general terms instead.

Unedited response types

Unedited responses to the Hispanic origin question were categorized into three major types: 1) no origin response, 2) single origin response, 3) multiple origin response (see figure 2). The formulation of these response types was based on the frequency and type of single and multiple responses to the Hispanic origin question found in the HCUF.

No origin response group

The first type of response group is composed of those individuals who did not report an origin in the Hispanic origin question. They neither marked a check box nor provided a response entry in the write-in line. The percent of individuals who left the Hispanic origin question blank is the item-nonresponse rate and is an indicator of how well the question was received and understood by the general population (For a more detailed discussion of non-response rates to the Hispanic origin question, see Cresce, Ramirez, and Spencer, 2001.)

Single origin response group

The second type of response group is comprised of those individuals who reported a single origin, either by checking a single check box or by writing in a single entry in the write in line of the Hispanic origin question. As instructed by the question itself, these individuals indicated that they were either Non-Hispanic origin or an Hispanic origin.[8]

Multiple origin response group

The third type of response group is comprised of individuals who reported more than one origin entry to the Hispanic origin question, regardless of whether they reported Non-Hispanic or general Hispanic terms. Three different kinds of multiple response groups were defined. Those individuals who reported two or more Hispanic groups such as Mexican and Cuban, or Puerto Rican and Dominican, but no non-Hispanic response, were categorized into the *Multiple Hispanic origin* group. Individuals who reported multiple Hispanic origins, including general terms such as Hispanic, Latino, Spanish or any other general term, were also included in this group. People who reported that they were both non-Hispanic and Hispanic were categorized into the "*Part*" *Hispanic origin* group.[9] These are individuals who identified as being both Non-Hispanic or Hispanic by either checking the non-Hispanic box and/or writing in a non-Hispanic entry such as French, German, or Portuguese etc, and either checking one of the Hispanic origin boxes and/or writing in a Hispanic entry such as Colombian, Argentinean, Latino, or Panamanian, etc.

---

[8] Individuals who reported only the term "American" in the Hispanic origin question were coded as Non-Hispanics.

[9] Because only one response to the Hispanic origin question was allowed (although multiple responses were captured), those individuals who reported both non-Hispanic and Hispanic terms were considered Part-Hispanics for editing purposes. Due to the nature of the edit rules and procedures, all of these individuals existed only in the (HCUF) and were later edited/allocated out into Hispanic or non-Hispanic in the creation of the (HDF).

The final multiple response group is comprised of individuals who reported only two or more Non-Hispanic terms and thus were placed into the *Multiple Non-Hispanic origin* group. Examples of how these individuals identified include two write-in entries such as French and German.

Figure 2. Topology of Origin Responses

1. No Origin Response
2. Single Origin Response
   a. Non-Hispanic
   b. Hispanic
3. Multiple origin responses
   a. Multiple Hispanic Origin
   b. Part Hispanic Origin
   c. Multiple Non-Hispanic Origin

Source: Census 2000

Original edit procedures for the Hispanic origin question

Data processing procedures for Census 2000 edited existing values and imputed values for missing data in the question on Hispanic origin. The edit and imputation process was divided into five basic parts: pre-editing, within-household imputation, hot deck imputation,[10] substitution, and group quarters editing (Cresce, Ramirez, and Spencer, 2001). The editing and imputation procedures for the Hispanic origin question at the Census Bureau reflect over 30 years of subject matter expertise. Because the Census Bureau has access to related origin information provided by the respondent on the questionnaire , such as Spanish surname or a Hispanic origin response in the question on race, editing and imputation for missing origin values have been greatly improved.

Pre-edit Procedures

Before any imputation was done on missing origin values, the unedited data underwent a series of edit procedures. The edit procedures for the Hispanic origin question consisted of the following rules:

[10] A hot deck is a data table (or "matrix") in which the values of reported responses, stratified by selected characteristics of the respondents (e.g., age, sex and race) were stored and updated on a flow basis and used as needed to assign values of the variable in question to people with similar characteristics who did not have a response.

1) Convert checked box marks into corresponding three-digit codes (e.g., if someone had checked the Mexican box "yes" then convert that check box into code 210);

2) Ensure that all write-in responses were valid and coded appropriately;

3) Override the code for the "Other Spanish/Hispanic/Latino" check box with the specific code for any origin that was provided, if any .[11]

In addition, multiple checkboxes that were marked were converted into one single output code unlike the race question, where only one single origin response was allowed in the Hispanic question. A single response was achieved as follows:

a) if all the multiple responses were Hispanic, then the respondent was assigned a code of (291) "Multiple Hispanic."

b) If all the multiple responses were non-Hispanic, then the respondent was assigned a Code of (190) "Multiple non-Hispanic"and,

c) If all the multiple responses were combinations of Hispanic and non-Hispanic terms (Part- Hispanics) then the responses were blanked and a single origin was imputed for the respondent using first within-household imputation or, failing that, hotdeck imputation.

This last step was undertaken because under current OMB standards, people must be classified as either Hispanic or not-Hispanic. For OMB purposes, both the Multiple-Hispanic and the Multiple-non-Hispanic responses can be categorized as Hispanic and non-Hispanic respectively. In Census 2000, approximately half of the Part-Hispanic responses were imputed as Hispanic and the other half were imputed as Non-Hispanic.[12] Based on internal edit reviews since Census 2000, several improvements to these pre-editing procedures have been proposed and are discussed later.

[11] For example,, the code for a write-in response of "Guatemalan" (code 222) replaced the check box code for "Other Spanish/Hispanic/Latino" (code 280).

[12] There were a total of about 700,000 Part Hispanics in Census 2000, (See Table 1 in this report). Imputation results were based on internal edit tally files from the Decennial Census 2000. For more information about Census imputation procedures, refer to Working Paper #65, "Evaluating Components of International Migration: Quality of Foreign-Born and Hispanic Population Data, by Arthur R. Cresce, Roberto R. Ramirez, and Gregory K. Spencer" at www.census.gov/population/www/documentation/twps0065.html.

Within-Household Imputation

If the respondent did not have a value for origin after pre-editing procedures, a value was imputed using a value from other household members in a particular sequence of household donor precedence. This precedence was based on household relationship. Thus, if members of a particular household were missing origin (donees), then they were assigned the origin of the householder (donor). On the other hand, if the householder was missing origin, then origin would be assigned to the householder based on the origin of the spouse of the householder. In addition, household members could only "donate" an origin if the household member needing an origin (donee) had the same race as the donor.

Hotdeck Imptuation

If an origin could not be imputed from other members of the household, an origin was imputed from another person of the same race and age group in a different household based on whether the person needing an origin had a Spanish surname.[13] People with a reported origin and a Spanish surname donated their origin to the Spanish surname-assisted hot deck. People with a reported origin and a non-Spanish surname donated their origin to the non-Spanish surname assisted hot deck. All other people with a reported origin donated their origin to a non-surname assisted hot deck. If a person requiring an origin from the hot deck has a Spanish surname, he or she would receive an origin from the Spanish-surname-assisted hot deck. If a person requiring an origin from the hot deck had a non-Spanish surname, he or she would receive an origin from the non-Spanish-surname-assisted hot deck. All other people requiring an origin from the hot deck would receive an origin from the non-surname-assisted hot deck. Census 2000 was the first decennial census to use surname-assisted hot decks.

Substitution

In some cases “substitution” (or “whole household substitution”) was used when no information was provided on the form. In this case, all the characteristics (including origin) of a nearby household of the same size were assigned (or substituted), using a substitution hot deck, into the household lacking the characteristics.

[13] A surname was determined to be “Spanish” if at least 10 people provided that surname and at least 85 percent of them reported as Hispanic origin at the state level. The same criteria were applied to see if a surname could be classified as not Hispanic. If a surname did not register 85% as either Hispanic or non-Hispanic at the state level, then it was classified as other (non-surname). For more information about hotdeck procedures, please refer to Working Paper #65, “Evaluating Components of International Migration: Quality of Foreign-Born and Hispanic Population Data, by Arthur R. Cresce, Roberto R. Ramirez, and Gregory K. Spencer” at www.census.gov/population/www/documentation/twps0065.html.

Group Quarters

A similar edit and imputation procedure to that for households was used for people in group quarters, except that separate hot decks for the group quarters population were stratified by age and six group quarters types: correctional institutions, nursing homes, college quarters, military quarters, other institutions, and all other group quarters.

Improvements to the Hispanic Origin Question Edit Procedures

Two areas of the original pre-edit procedures for the Hispanic origin question have been identified as contributing to the relatively large size of the Multiple origin population. The first area was the non-removal of general Hispanic terms (such as "Hispanic," "Spanish," and "Latino") when combined with a specific Hispanic terms. The race pre-edit procedure had a rule that when there was a combination of general and specific responses within the same race groups (for example, "Chinese and Asian"), the general response was dropped, leaving the single response "Chinese." This was not done in the Hispanic origin edit. Thus, instead of assigning a code of 291 (multiple Hispanic) to those people who reported "Mexican and Latino," the pre-edit procedure should have removed the general Hispanic term and reassigned this person a single origin response code of 211 (Mexican). There were 327,400 people nationally who reported a specific Hispanic term along with a general Hispanic term. If a pre-edit like the race pre-edit had been implemented during the final edit process, the total multiple Hispanic count would have been reduced by 54 percent, from 586,000 to 267,000 (see Table 1). This is a major reduction in the original count of the Multiple Hispanic origin count and caution should be taken when examining this group.

The second area identified after Census 2000 were cases where all the checkboxes on the Hispanic origin question were marked (excessive reporting) but not blanked for single origin imputation. These were individuals who had checked every box in the Hispanic origin question and may have also provided write-in responses. There were 67,485 such cases in 2000 (over twice as large as the Multiple Non-Hispanic population, see Table 1). Although this type of response is theoretically possible for some individuals of very diverse backgrounds, the Census Bureau thought it was highly unlikely for many individuals in the general population to report this way. The pre-edit procedure for the question on race had such a rule to blank cases where all the race checkboxes were marked. If this pre-edit had been implemented during Census 2000 (blanking these excessive responses), the total Part-Hispanic origin count would have been reduced by 8.7 percent, from about 775,000 to 707,000.

The combined effect of implementing both pre-edit rules would have been to reduce the original multiple origin count of 1.4 million by 386,000 to a new revised total of a little more than 1.0 million, a decline of 28 percent. For the purposes of this paper, we define the multiple origin population (1,001,344) as the cases that remain after applying the new improved pre-edit procedures (see Table 1).

Demographic Variables Used in the Analysis of multiple origin reporting

To get a better understanding of the sociodemographic characteristics of the multiple origin population in the United States and to determine which characteristic is the best predictor of multiple origin reporting (as measured by odds ratios), the following 100 percent items were used: age, sex, race, questionnaire form type, and tenure (see Tables 5 and 6). Table 5 shows the multiple origin population by type by these aforementioned variables and Table 6 shows multiple logistic regression results (odds ratios) by these same variables but with collapsed versions of the variables. The reported age of the respondent as of April 1, 2000 was coded into three broad age categories: (0-34 years, 35-64 years, and 65 and older) in Table 5. In the logistic model (see Table 6), age was collapsed into a dichotomous variable (1=0-34 years, 0=35 and older). Sex was coded in two groups: '1' for females and '0' for males for both descriptive purposes and for the logistic model. Race was coded into the following the major race groups: (White alone, Black alone, American Indian and Alaska Native alone, Asian alone, Native Hawaiian and Other Pacific Islander alone, Some other race alone, and Two or more races). For the logistic model and for sample size purposes, only two race categories were used: those individuals who only reported a single race (coded as '0') versus those who reported more than one race (coded as '1').

The Census Bureau used a number of different form types during Census 2000 to enumerate the population. For the purposes of this paper, form types were coded into two broad categories: Mail return forms (coded as '1') and Enumerator forms (coded as '0'). Mail return forms are those questionnaires that were sent out by mail and returned by mail. Mail return forms came in two types: "Mail Short Form (MSF) and the Mail Long Form (MLF)." Enumerator forms on the other hand are those questionnaires that the census worker used during the face to face enumeration process during the nonresponse followup phase of the census. Information about housing tenure was also collected. Owner-occupied housing units were coded as '0' while renter-occupied housing units were coded as '1'.

Multiple Logistic Analysis

To ensure sufficient sample size for the multiple logistic model, each of the demographic variables was coded into dichotomous variables. Given the categorical nature of the variables in this analysis, logistic regression analysis was used to estimate the log-odds of the outcome variable multiple origin reporting with the covariates of sex, age, race, form type, and tenure. Only the main effects model was examined. Logistic regression analysis allows for the examination of which demographic variables best predict multiple response after controlling for their effects simultaneously. Logistic regression parameter estimates and odds ratios for each of the covariates in the model were also produced for analytical purposes. Data processing, coding, descriptive statistical analysis, and logistic regression analysis were conducted using the Statistical Application Software Package (SAS). No statistical adjustments for sampling error (i.e., weights, design factors) were made in this analysis because data for the entire population were used and not sample data.

In logistic regression notation, the main effects model examined in this paper is expressed as follows: (1) logit p= a + b1 (Sex) + b2 (Age) + b3(race) + b4(Mode) + b5(Tenure) where the outcome variable multiple origin is a function of sex, age, multiple race, mode, and tenure. Figure 3 shows how each of the variables in the logistic model were coded.

Figure 3. Variables Used in the Multiple Logistic Model.

Variable	Values	Level
Sex	1	Female
	0	Male
Age	1	Under 35
	0	35 and over
Race	1	2 or more races
	0	Single race
Mode of data collection	1	Mail return form
	0	Enumerator form
Tenure	1	Renter
	0	Owner
Origin	1	2 or more origins
	0	No origin or single origin

Source: Hundred Percent Detailed File, Census 2000

Results and Key Findings

General Analysis

Table 2 shows that 94.3 percent of the total U.S. resident population provided a single ethnic origin response, while 5.4 percent left the Hispanic origin question entirely blank (nonreponse rate). Nationally, only 0.4 percent reported more than one ethnic origin. Similar results were observed by Census regions with the West region having the highest reported multiple origin responses (0.6 percent). Among the individual states, District of Columbia (10.7 percent), Hawaii (7.4 percent), Delaware (6.8 percent), Georgia (6.8 percent), Mississippi (6.8 percent) and Alabama (6.7 percent), had the highest non-response rates while Iowa (96.5 percent), Nebraska (96.5 percent), and North Dakota (96.4 percent) had the highest reported single origin responses. New Mexico (0.9 percent), California (0.8 percent), Hawaii (0.7 percent), and New York (0.6 percent) had the highest percentage of multiple origin responses.

According to Table 3, slightly over 1 million people in Census 2000 reported more than one origin. The most common multiple origin response type was Part-Hispanic origin, representing 70.6 percent of the total, followed by multiple Hispanic origin (26.7 percent) and multiple Non-Hispanic origin (2.7 percent). The West region had the highest reported Part-Hispanic origin responses (76.9 percent) while the Northeast had the lowest (57.2 percent). Multiple Hispanic origins were most commonly reported in the Northeast region (39.9 percent) and least likely to be reported in the West region (20.6 percent). The Northeast region also had the highest reported multiple non-Hispanic origins (2.9 percent) while the West region had the lowest (2.5 percent). Among the states, New Mexico, Wyoming, and Idaho had the highest percentage of Part-Hispanic origin responders (94.0 percent, 90.4 percent, and 89.7 percent, respectively) while Illinois (50.2 percent), New Jersey (44.9 percent), and Florida (44.1 percent) had the highest multiple Hispanic origin responses. Finally, Multiple non-Hispanic origin responses were most commonly reported in Hawaii, Rhode Island and Maine (19.9 percent, 8.6 percent, and 8.3 percent, respectively).

Table 4a shows the three largest multiple origin combinations among the total multiple origin population for the entire country. The largest multiple origin combination consisted of those individuals who reported they were both Non-Hispanic and Mexican. This specific combination represented 17.8 percent of the total multiple origin population. This finding is not surprising considering the Mexican origin population is the largest Hispanic group in the United States. The second largest multiple origin combination consisted of individuals who reported both Mexican and Puerto Rican origins. They represented 4.5 percent of the total multiple origin population. The third largest multiple origin group was composed of people who identified Non-Hispanic and Spanish. More specifically, these are individuals who marked the Non-Hispanic box and also wrote-in than they were Spanish in the write-in line. These top three multiple origin groups accounted for about a quarter (26.3 percent) of the entire multiple origin population (263,735 out of 1,001,344).

The three largest multiple Hispanic combinations are shown in Table 4b. For the multiple Hispanic origin population, the largest combinations were (Mexican and Puerto Rican) followed by (Puerto Rican and Cuban) and (Puerto Rican, Other Hispanic, and Dominican) (45,404; 20,947; and 9,888, respectively). These three groups alone represented more than one-fourth (28.5 percent) of the total Multiple Hispanic origin population (76,239 out of 267,105). The largest Part-Hispanic origin combinations in Table 4c were those individuals who reported (non-Hispanic and Mexican (178,428), followed by Non-Hispanic, Other Hispanic, and Spanish (39,903). The third largest Part-Hispanic origin combination was Non-Hispanic and Other Hispanic (33,941).[14] Over one-third (35.7 percent) of the total Part-Hispanic origin population was composed of these three combination groups (252,272 out of 707,070). Due to population size and confidentiality concerns, no specific multiple non-Hispanic combinations were shown.

[14] Other Hispanic in this case refers to the Other Hispanic check box only. There were no write-ins in this instance.

The percentage distributions of response type by sex, age, race, mode of data collection, and tenure are shown in Tables 5a and 5b. Over ten percent of people who identified as Black alone (10.5 percent) left the origin question blank. A similar percent was observed for Native Hawaiians and Pacific Islanders (10.8 percent), the highest among all the racial groups. People who were enumerated in the Census by enumerator forms were more than twice as likely not to have answered the origin question compared to people who responded by mail (10.5 percent vs. 3.4 percent, respectively). Finally, people who rented their homes were more likely than owners to have left the origin question blank (6.8 percent compared to 4.7 percent).

In general, there were no notable percent differences by sex for Multiple origin reporting. However, individuals under the age of 35, lived in renter-occupied housing units, and people who were enumerated by Mail forms were more likely to report Multiple origins (.5 percent, .6 percent, and .5 percent) than their counterparts. People who identified as Some Other Race or reported more than one race were especially likely to report multiple origins (1.5 percent and 2.2 percent, respectively; see Table 5a).

Among the three different types of Multiple origin reporting, Multiple Hispanic reporting was most common among individuals under the age of 35, compared to people aged 35 to 64, and 65 plus (34.4 percent, 13.1 percent, and 5.9 percent, respectively). In fact, all three age groups were mostly composed of Part-Hispanics with the 65 plus age group being the highest (91.2 percent; see Table 5b). Multiple Hispanic origin reporting was most common among the Some Other race alone population (44.3 percent), the highest among all race groups, followed by White alone (23.7 percent) and Native Hawaiian and Other Pacific Islander alone (19.9 percent). As with age, all the race groups were mostly composed of Part-Hispanics with American Indian and Alaskan Native being the highest (80.4 percent). On the other hand, multiple Non-Hispanic reporting was most common among Native Hawaiians and Pacific Islander alone and Asian alone (14.9 percent and 11.7 percent, respectively).

Multiple Hispanic origin reporting was more frequently observed in enumerator forms (43.2 percent) compared to mail questionnaires (23.4 percent). The opposite was true for people who reported Part-Hispanic, with mail questionnaires having a higher percent (73.9 percent) compared to enumerator forms (54.4 percent). No major difference was observed between the form types for multiple Non-Hispanic reporting. Renters were also more likely to report multiple Hispanic origin (30.8 percent vs 23.4 percent) while owners were more likely to report Part-Hispanic (74.0 percent vs. 66.4 percent). Similar results were observed for both owners and renters in the percentage of people reporting multiple Non-Hispanic origin.

Multivariate Analysis

The logistic regression parameter estimates are given in Table 6. The estimated log-odds of reporting a multiple origin was -6.9761 + -0.0273(sex) + .6026(age) + .08716(mode) + 1.7836(race) +.5325(tenure) when all the covariates in the model equaled 1. All of the covariates in the model were statistically significant at the .0001 level. Multiple race reporting had the highest effect on multiple origin with a log-odds parameter value of 1.7836. After controlling for the other covariates in the model, people under 35 were 1.8 times more likely to report a multiple origin than those people over 35 years. High odds ratios were also observed for those who answered by mail (2.4 times), provided more than one race response (5.9 times) and those who rented their homes (1.7 times) rather than owned one. Females were slightly less likely to report multiple origin than males (odds ratio=.97).

Part-Hispanic population

Given current standards set by OMB for the collection of data on ethnicity, respondents must be classified into one origin response category, either as Hispanic or Non-Hispanic for federal statistical reporting purposes. This standard still holds regardless of the number of origin terms a respondent provides on the questionnaire. For those respondents who reported more than one Hispanic or non-Hispanic term, the census edited them to either Multiple Hispanic (code of 291) or Multiple Non-Hispanic (code of 190) respectively (see Hispanic edit section). Although the multiple Hispanic responses posed no serious issue for OMB classification standards, they do affect the specificity of detailed Hispanic groups reported. So instead of assigning a code of 291 (Multiple Hispanic) for someone who had reported they were "Mexican and Latino" for example, they could have been coded to a single origin category as "Mexican" (code of 210) instead. In fact, most (327,400) of the multiple Hispanic responses given in Census 2000 were a specific Hispanic response combined with a general origin response (see Table 1).

How should Part-Hispanic responses be edited into a single origin response category? Because they must be resolved into either Hispanic or Non-Hispanic, significant research has gone into improving the origin edit to address this problem. The development of the two pre-edit procedures previously discussed were products of this research effort. In Census 2000, there were over 700,000 Part-Hispanics, and all their responses were ultimately blanked and allocated (approximately 50-50 proportionally) to a single response category (Hispanic or Non-Hispanic) via imputation (as previously discussed).

However, the question still remains: "To what extent were these people of mixed heritage?" In other words, did people of Part-Hispanic origins really possess such diverse backgrounds as they claimed on the questionnaire (being both Hispanic and a non-Hispanic). Although this question is difficult to answer without conducting reinterviews with the respondents to determine what they really meant by the way they reported, there are some ethnic-related questions on the long form that may shed some light on this issue. By examining the ethnic backgrounds of Part-Hispanics based on their responses to the ancestry, place of birth, and language use questions, a better picture of their "true" ethnic backgrounds may emerge.

Origin Index

The analysis in this section is divided into two parts. First, ancestry, place of birth, and language spoken in the home were examined by type of multiple origin, and second, these same ethnic indicators were combined into an origin index. The more Hispanic indicators one had, the more likely that person would be considered Hispanic for the purpose of this study. For example, a person who was born in Mexico, spoke Spanish at home, and had only Mexican ancestry would not be considered Part-Hispanic compared to someone who was born in the United States, spoke only English at home, and had French and Mexican ancestry. In addition, the origin index will also allow us to better understand the ethnic heritage of Part-Hispanics and help shed some light on whether or not people of mixed heritage were truly mixed as they claimed. These three ethnic variables were used because all three of them have at least 20 years of history with the Census and are the only ethnic-related questions on the long form.

Ancestry was first collected in the 1980 Census (sample). Up to two ancestry responses were collected and coded in Census 2000. Ancestry was coded into four broad groups; Not Reported, Hispanic only, Both Hispanic and Non-Hispanic, and Non-Hispanic only. Place of birth is one of the oldest questions in the Census, dating back to 1850. This question allows the identification of those respondents who were born in a Hispanic country, that is people born in countries in which Spanish is the primary language. There were three general places of birth coded: the United States, Hispanic countries, and other countries. The question on language spoken in the home was first used in the 1980 Census. This question allows for the identification of individuals who speak Spanish at home which provides another indicator of Hispanic origin. This language question was coded into the following three categories: English only, Spanish, and other languages.

Findings from the Origin Index

Table 7 shows the total multiple origin population by type tabulated by ancestry, place of birth, and language spoken at home. The results indicate that for the Part-Hispanic origin population (679,000), 37.7 percent reported having a single Hispanic origin ancestry while about a quarter of them (25.8 percent) reported having a mixed origin ancestry (Hispanic and non-Hispanic ancestry).[15] Over three fourths (77.1 percent) were born in the United States compared to only 18.4 percent born in a Hispanic country. Among those 5 years and older, 40.1 percent spoke Spanish at home and over half spoke English (55 percent). According to the origin index, 15.2 percent reported having Hispanic ancestry and spoke Spanish while only 1.5 percent reported having Hispanic ancestry and were born in a Hispanic country. Although about a third (33.5 percent) reported only having Hispanic ancestry, a similar percentage (29 percent) reported having no Hispanic indicators (that is, non-Hispanic background). Most Part-Hispanics, however

[15] The estimates in this section of the paper are based on responses from a sample of the population. As with all surveys, estimates may vary from the actual values because of sampling variation or other factors. All statements made in this section have undergone statistical testing and are significant at the 90-percent confidence level unless otherwise noted.

(about 70 percent), had either or a combination of Hispanic ancestry, spoke Spanish at home, and/or were born in a Hispanic country.

What does the origin index tell us about people of multiple origins? The origin index did an excellent job in discriminating the background of people who reported multiple non-Hispanic origins. For these particular individuals, for example, 91 percent of them also reported having had non-Hispanic origin indicators as shown in Table 7. For the most part, multiple non-Hispanics were of non-Hispanic Ancestry, born in a non-Hispanic country and spoke English or another language other than Spanish at home. The results for Multiple Hispanics and Part-Hispanics, on the other hand, were a little more mixed. According to the origin index, the vast majority of Multiple Hispanics (91.6 percent) had at least one Hispanic indicator. More specifically, 42.8 percent of them had one Hispanic indicator, 41.4 percent with two Hispanic indicators, and finally, 7.4 percent reported having three Hispanic indicators. Only 8.4 of Multiple Hispanics reported having no Hispanic indicators. This seems to suggest that these individuals had very strong Hispanic backgrounds. Nearly one-third of Part-Hispanics (29 percent) had no indicators of Hispanic ethnicity while most reported having at least one Hispanic indicator (71 percent). This finding seems to support the diverse background of these individuals.

Conclusion

In 2000, the Multiple Hispanic population represented only 0.5 percent of the total U.S. population (slightly over 1 million) and were mainly concentrated in the following states: California, Texas, Florida, and New York. The most common Multiple origin response type was Part-Hispanic, representing 70.6 percent of the total multiple origin population. People who were under age 35, identified with more than one race, and were enumerated by mail, were more likely than their counterparts to report a multiple origin. Finally, according to the Origin Index, most Part-Hispanics reported having some kind of Hispanic background as measured by Hispanic ancestry, Spanish language use, and place of birth. But nearly a third did not provide any of the indicators of Hispanic background. It appears that Part-Hispanics do indeed have diverse ethnic backgrounds as they claimed. The fact that 21.5 percent reported having a non-Hispanic ancestry and most spoke English only (55.0 percent) are strong indicators of people with multiple ethnic backgrounds.

The internal edit evaluation results show that if we had originally implemented the two new pre-edit procedures (i.e., the removal of general Hispanic terms and excessive origin reporting) to the Hispanic origin question during Census 2000, the total multiple origin population would have been reduced by nearly a third. This is a significant drop in the multiple origin population and caution should be used when describing the size and distribution of any given population, especially when edit specifications and procedures could have such an effect on the final size of a population.

Recommendations

Improvement of Origin Allocation: The use of Spanish Surname

One way to improve the edit procedures for Part-Hispanics is to use additional information from the questionnaire such as the person's surname to help determine if an individual should be coded as Hispanic or non-Hispanic. In Census 2000 for example, Spanish-surname-assisted hot decks were used to help assign missing origin when surname was available on the form. The use of auxiliary information from the questionnaire could more accurately allocate Part-Hispanics into the more appropriate single origin response category (Hispanic or Non-Hispanic). More research is needed to evaluate the impact this change could have on the origin edits that are currently being used to edit responses to the Hispanic origin question.

Self-Reported Multiple Origin v.s. Edit Created Multiple Origin

The allowance of multiple origin values (codes 291 and 190) to be assigned via imputation method (within household, hotdeck or substitution) for those individuals missing an origin response on the question on Hispanic origin artificially augmented the total multiple Hispanic and non-Hispanic counts by 7 percent and 10.8 percent, respectively (see table 8). Based on internal edit tally counts, for example, the final edited multiple Hispanic number of 627,676 was augmented by as much as 41,656 due to the fact that these additional multiple Hispanic codes (291) were added by means of imputation. For instance, of the 41,656 additional multiple codes added, 31,628 of these had been assigned from the within household edit (75.9 percent), 3,801 from hotdecks (surname and nonsurname; 9.1 percent) and 6,227 from substitution (14.9 percent). A similar pattern was observed for multiple non-Hispanics. The final edited multiple non-Hispanic number of 30,455 was augmented by as much as 3,283 due to imputation. Of the these additional multiple non-Hispanic codes (190) assigned, over 90 percent of them (3,069) had been assigned from the within household edit and the additional 214 from substitution (6.5 percent). One improvement to the origin edit would be not to allow these two codes to be imputed for missing origin values.

References

Cresce, A. and R. Ramirez (2003). *Analysis of General Hispanic Responses in Census 2000.* U.S. Census Bureau: Population Division Working Paper No. 72. Washington DC.

_____ R. Ramirez, and G. Spencer (2001). *Evaluating Components of International Migration: Quality of Foreign-Born and Hispanic Population Data.* U.S. Census Bureau: Population Division Working Paper No. 65. Washington DC.

Guzman, B. (2001). *The Hispanic Population.* U.S. Census Bureau: Census 2000 Brief. Washington DC.* [C2KBR/01-3]

Martin, B. (2002). *Questionnaire Effects on Reporting of Race and Hispanic Origin: Results of a Replication of the 1990 Mail Short Form in Census 2000.* U.S. Census Bureau: Alternative Questionnaire Experiment. Washington DC.

U.S. Census Bureau (1997). *Findings on Questions on Race and Hispanic Origin Tested in the 1996 National Content Survey.* Population Division Working Paper No. 16. Washington DC.

_____(1996). *Results of the 1996 Race and Ethnic Targeted Test.* Population Division Working Paper No.18. Washington DC.*

Roberto Ramirez
US Census Bureau, Population Division
Ethnic and Ancestry Statistics Branch
Federal Building Number 3, Room 2085, Washington, D.C. 20233
Roberto.R.Ramirez@census.gov