Processing of Race and Ethnicity in the 2004 National Survey on Drug Use and Health

E.A. Grau, P. Martin, P. Frechtel, J. Snodgrass, and R. Caspar
RTI International, Inc.
egrau@rti.org

Key Words: Race, Ethnicity, Editing, Imputation, Predictive Mean Neighborhoods

1. Introduction

The National Household Survey on Drug Use and Health (NSDUH), formerly known as the National Household Survey on Drug Abuse (NHSDA) prior to 2002[1], is the Federal Government's primary source of national data on the use of tobacco, alcohol, and illicit substances. The survey also contains questions on health, illegal behaviors, and other topics associated with substance use. Research Triangle Institute (RTI), working under contract to the Substance Abuse and Mental Health Services Administration (SAMHSA) is responsible for conducting sample selection, data collection, data processing, analysis, and reporting activities in the study. Since the inception of NSDUH, questions have been included to determine the race and ethnicity of each respondent. Race and ethnicity are routinely used as part of the demographic breakdowns in the analyses and the various reports generated from the survey. From 1971 to 1998, the race and ethnicity questions underwent few changes. However, along with the switch from paper-and-pencil interviewing (PAPI) methods of questionnaire administration to computer-assisted interviewing (CAI) methods in 1999, the race and ethnicity categories were updated pursuant to new Office of Management and Budget (OMB) directives. This report details how race and ethnicity data were recorded in the NSDUH since the 1999 CAI, as well as summarizing how these data were processed.

2. Office of Management and Budget (OMB) Directives on Race and Ethnicity

Prior to 1999, the instrument for the NSDUH included two questions to collect data on race and ethnicity in a manner consistent with OMB's 1977 Statistical Policy Directive No.15 (OMB, 1977). In keeping with this directive, the ethnicity question asked respondents whether they were Hispanic or of Spanish origin, and the race question asked respondents to select exactly one of the following five categories to describe themselves: white, black, American Indian or Alaska native, Asian or Pacific Islander (including Asian Indian), and other (specify). Directive No. 15 was developed to provide a common language to promote uniformity and comparability across Federal surveys for data on race and ethnicity. Development of these data standards stemmed in large measure from the need to enforce civil rights laws.

In response to criticisms that the minimum categories set forth in Directive No. 15 did not reflect the diversity of the U.S. population, OMB announced in July 1993 that it would undertake a comprehensive review of the categories used to collect data on race and ethnicity. OMB established the Interagency Committee for the Review of the Racial and Ethnic Standards in 1994. This committee worked with Federal agencies to develop recommendations for enhancing the accuracy of the data on race and ethnicity collected by the Federal Government. The committee's work included a request for public comment on the Directive No. 15 standards and research and testing to assess the possible effects of implementing the suggested changes on the quality and utility of the resulting data.

In October 1997, OMB released a notice, "Revisions to the Standards for the Classification of Federal Data on Race and Ethnicity" (OMB, 1997), which summarizes the results of both the research and the public comment. It also provides the new standards for maintaining, collecting, and presenting Federal data on race and ethnicity. The standard race question now includes five categories: American Indian or Alaska Native, Asian, black or African American, Native Hawaiian or Other Pacific Islander, and white. There are two categories for data on ethnicity: Hispanic or Latino and Not Hispanic or Latino. The OMB notice provides the following definitions for all of these categories:

American Indian or Alaska Native. A person having origins in any of the original peoples of North and South America (including Central America), and who maintains tribal affiliation or community attachment.

Asian. A person having origins in any of the original peoples of the Far East, Southeast Asia, or the Indian subcontinent including, for example, Cambodia,

[1] For the sake of clarity, the term NSDUH will be used throughout this document.

China, India, Japan, Korea, Malaysia, Pakistan, the Philippine Islands, Thailand, and Vietnam.

Black or African American. A person having origins in any of the black racial groups of Africa. Terms such as "Haitian" or "Negro" can be used in addition to "black or African American."

Hispanic or Latino. A person of Cuban, Mexican, Puerto Rican, Cuban, South or Central American, or other Spanish culture or origin, regardless of race. The term "Spanish origin" can be used in addition to "Hispanic or Latino."

Native Hawaiian or Other Pacific Islander. A person having origins in any of the original peoples of Hawaii, Guam, Samoa, or other Pacific Islands.

White. A person having origins in any of the original peoples of Europe, the Middle East, or North Africa.

The notice states that respondents must be offered the option of selecting one or more racial designations.1 The notice also suggests that to provide flexibility and ensure data quality, separate questions should be used wherever feasible for reporting race and ethnicity. When race and ethnicity are collected separately, ethnicity should be collected first.

Finally, the notice states that when aggregate data are presented, data producers should provide the number of respondents who selected only one category separately for each of the five racial categories. In addition, analysts are strongly encouraged to provide detailed distributions, including all possible combinations, of multiple responses for the race question. If data on multiple responses are collapsed, at a minimum the total number of respondents reporting more than one race should be provided as part of the data file.

OMB indicated that the new standard for collecting race and ethnicity superseded all previous requirements and for new or revised data collection activities, the standards were to be put into place by January 2003. The NSDUH instrument conversion from PAPI to CAI in 1999 afforded the opportunity to make this update. The new standards for collecting race and ethnicity were incorporated into the newly computerized 1999 NSDUH data collection instrument.

In October 2000, OMB approved the 2001 NSDUH with the conditional term that SAMHSA include a single best race (also referred to as "main race"), question for respondents who answered with more than one race, and that they report to OMB on the effects of this question. In particular, SAMHSA was to report to OMB on the interaction between Hispanic status, age, and responses to the race question. OMB used this information for research purposes in support of the Tabulation Working Group of the Interagency Committee for the Review of Standards for Data on Race and Ethnicity. The requested data was sent to OMB in April 2002. Starting with the 2003 NSDUH, the main race question was removed from the questionnaire.

3. Race Questions in Previous (Since 1999) and Current NSDUH Questionnaires

In keeping with the new standard, the NSDUH instrument included two questions to collect race and ethnicity. The ethnicity item was read by the interviewer who then entered the respondent's answer into the computer-assisted personal interviewing (CAPI) program. The question (QD03) asked:

> *Are you of Hispanic, Latino, or Spanish origin or descent?*

It should be noted that OMB's 1997 notice indicates that use of the phrase "Spanish origin" may also be included in the question text if desired. This question remained unchanged in each of the six survey years from 1999 to 2004.

The race question (QD05) was also included in the CAPI module in all survey years from 1999 to 2004. When administering this question in 1999, the interviewer handed the respondent a card with the following categories listed: white, black/African American, American Indian or Alaska native, Native Hawaiian, Other Pacific islander, Chinese, Filipino, Japanese, Asian Indian, Korean, Vietnamese, and Other Asian.

The question was worded as follows:

> *Which of these groups describes you? Just give me the number or numbers from the card.*

The respondent could select one or more of the listed categories from the card, or he or she could indicate that none of the categories applied, whereupon the interviewer could write in a response.

Although this question included more categories than explicitly included in OMB's 1997 notice, the categorization could have been collapsed to provide the level of detail required by OMB. The 1997 notice explicitly states that additional detail is allowable so long as the listing allows for back coding to the original six categories defined in the notice.

These categories were maintained in subsequent years, though the information was obtained

differently. Most notably, the categories were revised to align more directly with OMB's 1997 notice. Additional information on specific Asian races that had been collected within the initial race question in 1999 was moved to a follow-up question. Also, category for American Indian or Alaska native was revised to provide an explicit definition of "American Indian."

In surveys after the 1999 one, the categories listed in the first race question were: white, black/African American, American Indian or Alaska native (American Indian includes North American, Central American, and South American Indians), Native Hawaiian, Other Pacific islander, Asian (for example: Asian Indian, Chinese, Filipino, Japanese, Korean, and Vietnamese), Other (specify).

For respondents who selected the "Asian" category, an additional question (QD05ASIA) was included:

> *Which of these Asian groups best describes you? Just give me the number or numbers from the card.*

The following Asian categories were listed: Asian Indian, Chinese, Filipino, Japanese, Korean, Vietnamese, Other (specify)

In the survey years from 1999 to 2002, a follow-up item was also included in the instrument for respondents who selected more than one race. This item, referred to as the "main race question" earlier, was included to ensure categories could be back coded to the original mutually exclusive race categories. The item (QD06) asked:

> *Which **one** of these groups, that is [LIST RACES SELECTED FROM QUESTION ABOVE], **best** describes you? SELECT ONLY **ONE** ANSWER.*

The same race categories included in QD05 above were used for this item. However, only the categories that had originally been entered by the interviewer were available to be selected. For example, if a respondent had originally selected white and Korean, then Japanese could not be selected as the best descriptor in the follow-up item.

With two questions about race in the NSDUH surveys from 2000 onwards, a reprogramming of the CAI questionnaire was required so that multiple races reported in either of the two race items—the original question and the follow-up item to determine the specific Asian group—would trigger this main race question. In the 2001 and 2002 surveys, an additional category was added to the "main race" question for respondents who reported that **none** of the races they had reported was the best descriptor of their race. This main race question (QD06) was eliminated in the 2003 survey year, following an OMB directive subsequent to the 1997 directive that required all agencies to stop collecting data on main race. This did not have an impact on the processing of tables involving race and ethnicity, since all these tables included a level for "more than one race". However, models used in subsequent processing for imputation and small area estimation used a race variable that assumed a main race. Strategies used to deal with the elimination of this question for these subsequent processes are discussed in Section 4.

4. Editing of Race Variables

In this section, the methods used to process the data from the race and ethnicity questions are summarized. In keeping with OMB guidelines, "Hispanic/Latino" is considered an ethnicity, not a race. However, many respondents identified their race as "Hispanic/Latino" or indicated a Hispanic group in the other-specify response, resulting in a considerable amount of missing data for the race question. For this reason, drug use patterns were tabulated with race and ethnicity considered together as cross-classifying variables (in effect, treating "Hispanic/Latino" as a race).

As a result of the confusion between Hispanicity and race, Hispanicity was used in the editing of race and vice versa. In the process of editing race, the other-specify response to the Hispanic group question (QD04) was consulted (if it existed) if no race information was identified in QD05 or QD05ASIA. Similarly, in the process of editing the Hispanic group, the other-specify responses to the race questions (QD05 and QD05ASIA) were consulted (if they existed), if no Hispanic group information was identified in QD04. The identification of Hispanic groups is not discussed in this document.

In summary, the only editing that occurred with the race variables involved the integration of the other-specify information with the given categories of the questionnaire. In this section, the methods used to codify and categorize the other-specify information is described first, followed by the description of the approaches to presenting the race information.

4.1 Classification of Race Other-Specify Codes

In general, respondents were able to identify a race for themselves with one or more of the given categories in the vast majority of cases (ranging between 96.0% and 96.6% in the NSDUH surveys between 1999 and 2004). However, in some cases respondents either indicated that none of the given categories applied to them and they wrote in a response (an "other-specify" response), or they combined a given category with a write-in response.

All other-specify responses from QD04, QD05, and QD05ASIA were assigned both a race code and a Hispanic code. Each of the race codes was mapped to at least one of the categories described in Section 3, or to some other code that was informative in the final imputation described in Section 5. A summary of categories of other-specify codes and how they were handled is given below.

Race codes that were assigned according to other-specify responses were of four types: (1) directly mapped codes; (2) indirectly mapped codes (these required a quick imputation using a randomly generated number); (3) codes informative for formal imputation procedures; and (4) noninformative codes. The procedures applied to directly mapped codes and indirectly mapped codes resulted in values that were considered "final." The two other types of codes resulted in incomplete values requiring imputation, and were either informative or noninformative for the formal imputation procedures as described in Section 5. Each of the four types of codes is discussed below.

4.1.1 Directly Mapped Codes

The directly mapped codes were mapped to one or more of the categories given in the questionnaire (see Section 3). There were two types of directly mapped codes: a) racial category codes, and b) geographic category codes. Racial category codes were exactly equivalent to one or more categories in QD05 or QD05ASIA, and mapped directly to those categories regardless of whether the write-in response was in QD05 or QD05ASIA. (Respondents were still considered at least part Asian even if the write-in response in QD05ASIA was non-Asian. The racial compositions of respondents who entered a non-Asian racial category in QD05ASIA were determined on a case-by-base basis.) For example, a response such as "Han" mapped directly to a category in QD05ASIA ("Chinese") and the response "mestizo" mapped directly to two categories in QD05, "white" and "Native American." Geographic category codes corresponded to a country where census data indicated a racially homogeneous society. The mapping of geographic category codes for non-Asian countries depended upon whether the write-in response was in QD05 or QD05ASIA. For example, an entry of "Polish" in the QD05 other-specify mapped to white, since the Polish census data indicated that nearly all Poles were white. An entry of "Polish" in the QD05ASIA other-specify mapped to "other Asian." Geographic category codes also included ethnic groups where the racial identification was not immediately obvious. For example, a response of "Arab" would be automatically mapped to "white" if the response was a write-in answer for QD05. However, as with the "Polish" entry, if the "Arab" response was a write-in response in QD05ASIA, the respondent was considered “other Asian”.

4.1.2 Indirectly Mapped Codes

Codes that were indirectly mapped also corresponded to countries where census data were used, but for indirect mapping the countries were racially heterogeneous. A racial category was chosen by generating a random number and allocating the race based on a comparison of the random number with the proportions of races in the country's census. For example, an entry of "Bolivian" would have a 55 percent chance of being allocated to the American Indian category, since the latest Bolivian census indicated 55 percent of Bolivians were American Indian. For countries where the census indicated a small proportion of some non-distinct category such as "other", and the randomly generated number indicated an allocation to this proportion was called for, the final race was left to imputation. If two or three heterogeneous countries were entered in the other-specify response (e.g., "Bolivian and Peruvian"), the final race was allocated using the following procedure: (1) randomly assign races based on the proportions for each country mentioned; (2) combine the results. Exceptions to these rules occurred with the Hispanic Group categories Mexican, Puerto Rican, Cuban, Dominican, Spanish (from Spain), and Central or South American (no country given), which were given codes described under the next subheading, with a final value determined using the formal imputation procedures described in Section 5.

4.1.3 Codes Informative for Formal Imputation Procedures

Definitive information about the respondent's race. However, the responses were used to limit the final imputation described in Section 5. For example, a response of "mixed" resulted in an imputation among donors with two or more races, and a response of "brown" resulted in an imputation among donors who were not single-race white.

4.1.4 Noninformative Codes

Finally, a noninformative response (e.g., "American") that was not accompanied by a response to one of the given (non-other-specify) categories resulted in an unrestricted imputation. Religious identifications (e.g., "Muslim") were considered noninformative, even if the religion was usually associated with a particular ethnic group (e.g. "Shinto" is usually associated with Japanese).

4.2 Subsequent Editing of Race Other-Specify Codes

Subsequent to the initial mapping of the other-specify codes, edits were sometimes implemented that revised or clarified the initial mapping before final races were allocated. These edits were necessary if multiple sources of information, including other-specify responses, provided conflicting or confusing information. These edits were implemented when (1) the final mapping depended upon the source question; (2) responses were given to both the other-specify and non-other-specify categories of QD05 or QD05ASIA; or (3) different other-specify responses were present in at least two of QD04, QD05, and QD05ASIA. In some cases, it was necessary to individually examine the responses in order to determine the appropriate mapping. Each of these is discussed below.

4.2.1 The Final Mapping Depended upon the Source Question

In some cases, the final mapped value depended upon whether the other-specify code was in QD04, QD05, or QD05ASIA. An example from directly mapped codes is "Indian." This response would have been mapped to "American Indian" if the other-specify response was in QD05, but "Asian Indian" if the other-specify response was in QD05ASIA. Indirectly mapped codes could also have depended upon the source question. The census data from many countries included Asian categories. If the other-specify response was in QD05ASIA, the random imputation to a census category was limited to the Asian categories. Other-specify responses that were not specifically Asian sometimes occurred in the other-specify of QD05ASIA, as noted previously. These were carefully examined, but the "Asian" part of the response was always preserved.

4.2.2 Responses Given to Both Other-Specify and Non-Other-Specify Categories

If other-specify responses to QD05 or QD05ASIA accompanied responses to the given (non-other-specify) categories of QD05 and QD05ASIA, it was necessary to reconcile these responses. In some cases, the combination of responses mapped to one of the multiple race categories. There were instances, however, when the other-specify response was ignored because of responses to the non-other-specify categories. In particular, the other-specify response was always ignored if a non-other-specify category was selected, and the other-specify response was a geographic category code.[2] For example, if the interviewer selected the category for "black" for the respondent and also wrote in "Polish," it was assumed that the respondent was a black Pole, and for racial identification purposes, was considered single-race black. This was true even though the Polish census did not identify significant numbers of nonwhite peoples in the Polish population.

4.2.3 Different Other-Specify Responses Present in at Least Two of QD04, QD05, and QD05ASIA

In some instances, it was necessary to reconcile the other-specify responses to QD04, QD05, and QD05ASIA. In these cases, the responses were examined on an individual basis, and sometimes a new code was assigned that more accurately reflected the situation.

4.3 Approaches to Presenting the Race Information

Four different ways of presenting the race information resulted in four types of edited race variables: (1) individual indicator variables for each race; (2) broad racial categories, with descriptive levels for multiple race respondents, and levels with incomplete information useful for imputation; (3) broad categories with no multiple race information; and (4) detailed racial categories, with broad non-descriptive categories for multiple race respondents.

4.3.1 Individual Race Indicator Variables

Edited indicator variables, named EDQD05*xx*, were created which correspond to each of the 12 race categories described in Section 2. The "*xx*" represented a number between 1 and 12, each associated with a race category: (1) white; (2) black/African American; (3) American Indian or Alaska native; (4) Native Hawaiian; (5) Other Pacific islander; (6) Chinese; (7) Filipino; (8) Japanese; (9) Asian Indian; (10) Korean; (11) Vietnamese; and (12) Other Asian. For each of these variables, a nonzero value indicated that the respondent belonged to the group in question. A 13th indicator variable was also created (EDQD0513), which was a little different from the others. In particular, there was no specific level of QD05 or QD05ASIA which corresponded to it. It was used mainly to preserve a response of "Asian" to QD05, even if the respondent selected nothing in QD05ASIA.

---

[2] Actually, this "edit" was not "subsequent" to the initial mapping. Instead, the initial mapping was ignored under the circumstances described.

4.3.2 Broad Categories of Race, Descriptive Levels for Multiple Race, and Incomplete Information

The variable EDRACE summarized which of four broad race categories (white, black/African American, American Indian/Alaska Native, Asian/Pacific Islander) were identified in QD04, QD05, and QD05ASIA, in addition to any combination of those levels; it also had levels to indicate how the imputation should have been restricted based on the race of the donor. The first three broad race categories corresponded to EDQD051, EDQD052, and EDQD053 respectively; "Asian/Pacific Islander" was considered to have been identified if any of EDQD054-EDQD0513 was nonmissing. In all, EDRACE had 20 levels, 15 corresponding to the four race categories listed above and any combination of those categories. The 5 remaining categories were:

16 (multiple race, no other information), if an other-specify answer such as "biracial" or "mixed" appeared in QD04, QD05, or QD05ASIA
17 (nonwhite, no other information), if an other-specify answer such as "brown," "tan," or similar answers in Spanish appeared in QD04, QD05, or QD05ASIA
18 (white, or both white and American Indian/Alaska Native), if the random assignment of a census data code resulted in imputation restricted to donors who were either white, or both white and American Indian/Alaska Native
19 (not American Indian/Alaska Native, in part or in full), if the random assignment of a census data code resulted in imputation restricted to donors who were not American Indian/Alaska Native, in part or in full
20 (non-Hispanic Mexican), if "Mexican" was mentioned in the QD05 and/or QD05ASIA other-specify responses, but QD03 indicated "not Hispanic"

4.3.3 Broad Categories of Race, No Multiple Race

Because of the paucity and heterogeneity of multiple-race respondents, imputation models for race did not include a category for more than one race. Instead, predicted means were determined in multinomial logistic models with the same four categories identified in the previous section as “broad race categories”: white, black/African American, American Indian/Alaska native, and Asian/Pacific Islander. An edited variable was created that included the four broad categories given above. This variable was called EDRACEFORMODEL. Respondents who were missing values for EDRACE or had values of EDRACE between 16 and 20 had missing values for EDRACEFORMODEL.

In the survey years from 1999 through 2002, multiple-race respondents were assigned a single race based on the response to QD06, the multiple-race respondent's "main race." The handful of multiple-race respondents who did not answer QD06 were allocated a "main race" based on an arbitrary priority rule (black, Asian/Pacific Islander, American Indian/Alaska native, white). Imputation donors were chosen with predicted means for these four categories close to those of the recipient with missing race. A respondent was imputed as being more than one race if the selected donor also identified more than one race. However, in the 2003 and 2004 survey years, the "main race" question (QD06) was not included in the questionnaire. The respondent therefore did not have an opportunity to indicate a "main race", and assigning a main race using the priority rule for all multiple race respondents was not advisable, so a main race had to be assigned probabilistically using models.

Using data pooled across the survey years 2000-2002, a single race was imputed for multiple-race respondents using a series of logistic models. Eleven predictive mean models were fit, one for each multiple race category (EDRACE between the values of 5 and 15 inclusive). The parameter estimates from the models were used to impute a "main" or "best" race by the following procedure:

Step 1: Estimate the probability that each respondent would have mentioned each of the broad race categories indicated as their "main" race, using the coefficients from the appropriate predictive mean model.
Step 2: Randomly select one of the broad race categories based on these probabilities.

For example, consider a respondent in the 2004 NSDUH with EDRACE = 5 (white and black only). The covariates included in the model for respondents with EDRACE = 5 were age, region, race of householder, percentage of owner-occupied households, percentage Asian population, percentage American Indian population, and percentage black population. Using the values for these covariates for the 2004 respondent and the parameter estimates from the model, the probability that the respondent would have selected white as his main race could have been estimated. If this probability was estimated at 50 percent, a random imputation was done such that the respondent was assigned white as his main race with probability 50 percent and black as his main race with probability 50 percent.

4.3.4 Finer Categories of Race

EDNWRACE was a 16-level edited variable that included the same information as the 13 categories

identified in the individual race indicator variables (EDQD05xx), provided only one of the variables had a nonzero value. The following three additional categories were also created in situations where more than one of the 13 categories was identified:

(1) Native Hawaiian and Other Pacific Islander only, if both EDQD054 and EDQD055 were nonmissing, and all other EDQD05*xx* variables were missing

(2) Asian multiple category, if all of EDQD051-EDQD055 were missing (i.e., at least two of the ordinary Asian categories were selected)

(3) More than one of the broad categories of race

5. Race Variables: Imputation

As the editing of race and ethnicity were closely related in the NSDUH, imputation of race and ethnicity were also closely related. Race was used in the imputation of Hispanic origin and Hispanicity was used in the imputation of race. The method used for imputing missing values in the race variables is called the Predictive Mean Neighborhood method (PMN), which combines Predictive Mean Matching (Rubin, 1986) and Nearest Neighbor Hot Deck, described in Singh, Grau, and Folsom (2001). A neighborhood of potential donors is selected based on the predictive mean vector obtained from a multinomial logistic model. Even if a single variable was being imputed, the method was termed multivariate (MPMN) due to the multivariate nature of the predictive mean vector.

Table 1. Edited Race Variables and their Imputation-Revised Counterparts

Edited Race Variable	Imputation-Revised Race Variable
EDQD051	IRRACEWH
EDQD052	IRRACEBL
EDQD053	IRRACENA
EDQD054	IRRACENH
EDQD055	IRRACEPI
EDQD056-EDQD0513 (collapsed)	IRRACEAS
EDRACE	IRDETAILEDRACE
EDRACEFORMODEL	IRRACE2
EDNWRACE	IRNWRACE

Nearly all the edited variables had imputation-revised counterparts, as shown in Table 1 (some of the individual race category variables were collapsed at the imputation stage). For each of the imputation-revised variables, levels of the edited variable that indicated missing data or incomplete data were replaced with imputed values. The steps for model-building in the PMN procedure are noted below: setup for model building, computation of predicted means, assignment of imputed values, and constraints used on imputation neighborhoods.

5.1 Setup for Model Building

Each of the predictive mean models used survey weights to account for the survey design. These weights were equivalent to the sample design weights, but were adjusted to account for nonresponse at the household level using a simple ratio adjustment. The weights were also adjusted for item nonresponse using a weighting calibration model called the Generalized Exponential Model (Folsom and Singh, 2000). An interview respondent was considered an item nonrespondent for race if either EDRACEFORMODEL was missing, EDNWRACE was missing or was Asian with no specific Asian category given, or both. (If the respondent had missing data for either EDRACEFORMODEL or EDNWRACE, he or she also had missing data for the other edited variables in Table 1 [EDQD051-EDQD0513 and EDRACE].) The weights of the item nonrespondents were recalibrated among the item respondents using an exponential model. The covariates in these models included census region, household type (from the screener), centered age, percentage Hispanic population, percentage of owner-occupied households, percentage black population, percentage American Indian population, and percentage Asian population.

5.2 Computation of Predicted Means

Using the adjusted weights, the probability of selecting each race category was modeled within each age group using polytomous logistic regression.[3] The predictors included in the models were the same as those used in the item response propensity model for race.

The PMN method for race was multivariate, as opposed to univariate, because the predictive mean vector contained more than one element. The three elements in the vector were the predicted probability of being identified with each of the first three race categories (white, black, American Indian/Alaska native). The probability of being classified as

[3] SAS®-callable SUDAAN® was used to fit the polytomous logistic regression models. Details about the polytomous logistic regression model and additional references can be found in the *SUDAAN® User's Manual, Release 9.0* (RTI, 2004). SAS® software is a registered trademark of SAS Institute, Inc.; SUDAAN® is a registered trademark of RTI International.

Asian/Pacific Islander was not included because it was completely defined by the first three elements in the predictive mean vector, being calculated as one minus their sum. A predictive mean vector of three predicted means was created from the polytomous logistic regression model.

Conditional probabilities were calculated for the few item nonrespondents with EDRACE values of 18 or 19.

5.3 Assignment of Imputed Values

The predicted means were used to select a donor in a nearest neighbor hot deck. This donor was selected randomly from a pool of donors in a neighborhood that was defined by comparing the predicted means of the potential donors and the recipient. Because multiple variables were considered in the distance measure, "similarity" was defined in terms of the smallest Mahalanobis distance.

If the recipient had values that were missing for several of the variables listed in Table 1, the donor gave values for all relevant variables to the recipient. In most cases, this ensured consistency between each of the imputation-revised variables. An exception occurred when a respondent listed only one specific category of race, but indicated that he or she was of more than one race in the other-specify entry. In these rare cases, the respondent was "more than one race" in IRNWRACE, but only one race was given in the IRRACE*xx* and IRDETAILEDRACE variables.

5.4 Constraints on Neighborhoods

Constraints were placed upon the neighborhood based upon the information that was available about the item nonrespondent. These constraints were referred to as "logical constraints" if their absence potentially rendered the imputed value inconsistent with preexisting nonmissing values. For example, a response of "multiple race" in the other-specify response meant that donors had to have provided more than one race. Other examples included responses like "nonwhite," "brown," and "white or mestizo."

Other constraints, called likeness constraints, were placed upon the neighborhood to ensure that the donor and recipient were as alike as possible. The predicted means themselves were used as likeness constraints: each predicted mean from the recipient's model had to be within 5% ("delta") of the respective predicted means of each donor. (If the predicted mean was greater than 0.5, the percentage was calculated using the difference between the predicted mean and 1.) Other likeness constraints were used for variables that could not be included in the predictive mean model. These included requirements that donors and recipients (1) live in the same geographical primary sampling unit (called a "segment"); and (2) belong to the same Hispanic group, if applicable. As noted in Section 4, the latter constraint was applied if the respondent was Mexican, Puerto Rican, Cuban, Dominican, Spanish (from Spain), and/or Central or South American (no country given). If insufficient donors were available in the resulting neighborhood, the segment constraint on the potential donor was removed first. If no potential donors met the "delta constraint," the delta constraint was also removed. The Hispanic group constraints were never removed.

6. Imputation and Editing Summary for Race

To differentiate the final imputed values from nonmissing values, variables accompanying the imputation-revised variables gave information about the source of the information for the variable of interest. Through the survey years, these indicators showed that between 2 and 3 percent of all respondents required imputation, the vast majority of which were imputed within a neighborhood restricted by Hispanic group, and only a tiny handful were imputed within a neighborhood unrestricted by Hispanic group.

References

Folsom, R.E., and Singh, A.C. (2000) A generalized exponential model for sampling weight calibration for extreme weights, nonresponse, and poststratification, *ASA Proc. Surv. Res. Meth. Sec.* pp. 598-603.

Office of Management and Budget. (1977). *Race and ethnic standards for federal statistics and administrative reporting* (Statistical Policy Directive No. 15). Washington, DC: Author.

Office of Management and Budget. (1997). *Revisions to the standards for the classification of federal data on race and ethnicity.* Federal Register, 62 (210).

RTI. (2004). *SUDAAN user's manual*: *Release 9.0*. Research Triangle Park, NC: Author

Rubin, D.B. (1986) Statistical matching using file concatenation with adjusted weights and multiple imputations. J. Business Econ. Statist. **4**, 87-94.

Singh, A.C., Grau, E.A., and Folsom, R.E. (2001) Predictive mean neighborhood imputation with application to the person-pair data of the NHSDA." In *ASA Proc. Surv. Res. Meth. Sec.* [Available as a PDF at [www.amstat.org/sections/SRMS/proceedings]