Race/Ethnicity in the UW-BHS

Table of Contents:

I. Introduction

The primary aim of the UW-BHS is to explain post-secondary outcomes--to answer questions about who makes it to college, who finishes, how long it takes, and why certain students end up better off than others. While it's common knowledge that subpopulations, such as those defined by race/ethnicity, often experience different outcomes, few studies are designed to focus on how these subpopulations are defined and measured.

The UW-BHS focuses explicitly on the theory and measurement of race/ethnic subpopulations at every level of the research process, from study design and implementation to data processing and analysis. The study combines an array of self reported race/ethnicity measures with multiple supplemental data sources to obtain comprehensive, multidimensional accounts of each subject's race/ethnic identity. These highly detailed accounts are then simplified using an integrated coding typology that provides a straightforward yet flexible approach to reducing common sources of error in race/ethnic measurement. The results of these efforts provide an elegant and empirically defensible methodology for collecting and summarizing complex and often messy data on race/ethnic identities in social research.

IA. Background

In the U.S., standards for the measurement and classification of data on race and ethnicity are determined by the Executive Office of

Management and Budget (OMB), which specifies six irreducible categories that must be included in any statutory measure of race/ethnicity collected by the government and its contractors. They are:

-White
-Black
-American Indian and Alaska Native (AIAN)
-Asian
-Native Hawaiian and other Pacific Islander (NHOPI)
-Hispanic

In addition to specifying which groups must be included, the standard also specifies *how* each group should be measured, though practitioners have some flexibility on the latter.

Three facets of racial measurement are largely non-negotiable. First, race/ethnic identification must be self-reported unless it is impractical to do so (e.g., birth and death records). Second, only five of the six categories are classified as "racial" groups. Hispanics are defined as an ethnic group, members of which can be of any race. Thus, the OMB standards stipulate that race and Hispanic origin are distinct concepts, and must be measured separately whenever self-identification is used. Third, the standards stipulate that individuals may identify as many races as they wish.

Each of these requirements presents challenges. First, the decision to rely on self-reported race is a matter of necessity, since the census and many household surveys are conducted using self-administered mail-in questionnaires. Still, self-reports are potentially at odds with the way racial identification (and discrimination) takes place in daily life, certainly in instances of racial profiling, hate crimes, etc., but also in more mundane settings (e.g., coroner classification of the race of the deceased). If self-reported identities do not correspond to external appraisals, standard estimates of the size and characteristics of race/ethnic groups may be suspect.

Second, the distinction between race and Hispanic origin is problematic, since there is evidence that Hispanics often view themselves as a racial group. The end result is a typically high number of errant and missing responses to the official "race" question by persons of Hispanic origin. Listing Hispanic among the racial categories alleviates this problem, though this so called "combined format"[1] can only be used in instances in which self-identification is unfeasible (e.g., coroner classification).

The provision for multiple race responses was added during the mid 90s revision of the OMB standards. In addition to creating a large number of new racial categories and combinations, these changes created new problems regarding how individuals who report multiple races should be

[1] "OMB accepts the following recommendations concerning a combined race and Hispanic ethnicity question: When self-identification is used, the two question format should be used…when self-identification is not feasible or appropriate, a combined question can be used and should include a separate Hispanic category co-equal with the other categories."
(http://www.whitehouse.gov/omb/rewrite/fedreg/ombdir15.html)

counted for statutory purposes that require only one race or ethnicity. Strangely, while individuals can now identify multiple races, there are no provisions to identify multiple Hispanic origins (e.g., Mexican and Puerto Rican heritage), or to identify as Hispanic and non-Hispanic simultaneously.

Other challenges in the area of racial measurement lie not with the OMB standard but with how inquiries on race/ethnicity are implemented by researchers. First, OMB specifies only the *minimum* number of categories that must be distinctly measured. While these categories cannot be collapsed (e.g., Black and Asian cannot be combined), they can be subdivided *ad infinitum* (Chinese, Japanese, Cherokee, Mexican, etc), leaving researchers with a choice to limit measures to OMB groups or to measure race/ethnic diversity in greater detail.

Each option has drawbacks. On the one hand, pooling diverse origin groups into umbrella categories like Hispanic and Asian may confuse or offend respondents who identify specific national or ethnic identities. Pooling also combines under a common heading groups that might well differ in dramatic ways (Japanese vs. Cambodian, e.g.) Thus, questionnaires often measure detailed tribal, national, and regional origins by listing additional subcategories and/or providing space for "write-in" responses. On the other hand, the inclusion of detailed subgroups can dramatically expand the size of race/ethnicity queries (Census has codes for over 1000 ethnic groups), increasing the number categories and combinations by several fold.

Regardless of how many groups are listed, there will always be some individuals who refuse to identify with any of them. Thus, while the OMB standard is inclusive in theory, most implementations (including that used by the Census and many household surveys) include a residual "some other race" (SOR) category (typically accompanied by a write-in space) for those who feel their identities are not reflected in the OMB classification. While a portion of those who supply a write-in under SOR can be re-coded to an OMB category (e.g. Irish-American = White), other responses (e.g., "American") are less straightforward.

The UW-BHS includes measures and procedures to resolve each of the shortcomings listed above. Rather than limiting respondents to a preset number of vague racial or pan-ethnic categories, it includes multiple closed and open-ended inquires on race/ethnic background. Rather than leaving an intractable number of multiple race (or race by Hispanic origin) combinations, it utilizes a "primary identity" measure to simplify the complex identities of many Americans. Rather than relying solely on self-reports of race/ethnicity that have long dominated social research, the UW-BHS distinguishes population groups using an array of first- and third- person measures gathered from multiple perspectives and data sources.

These efforts begin with a baseline survey that measures race/ethnicity at a level of detail rarely used in empirical research. In addition to standard Census measures of race and Hispanic origin, the survey contains open-ended questions on ancestry, parental ancestry, primary identity, and reflected identity. Complementing this rich assortment of self-reported measures are supplemental data from multiple sources and measurement perspectives, including administrative records, a parental questionnaire, and third-person ("observed" race) appraisals from high

school yearbooks. These data are processed using a unified coding scheme that summarizes race/ethnic identities at various levels of detail while minimizing common sources of measurement error.

In the remainder of this section, I discuss each of these measurement sources in detail, starting with the design of the UW-BHS senior survey itself.

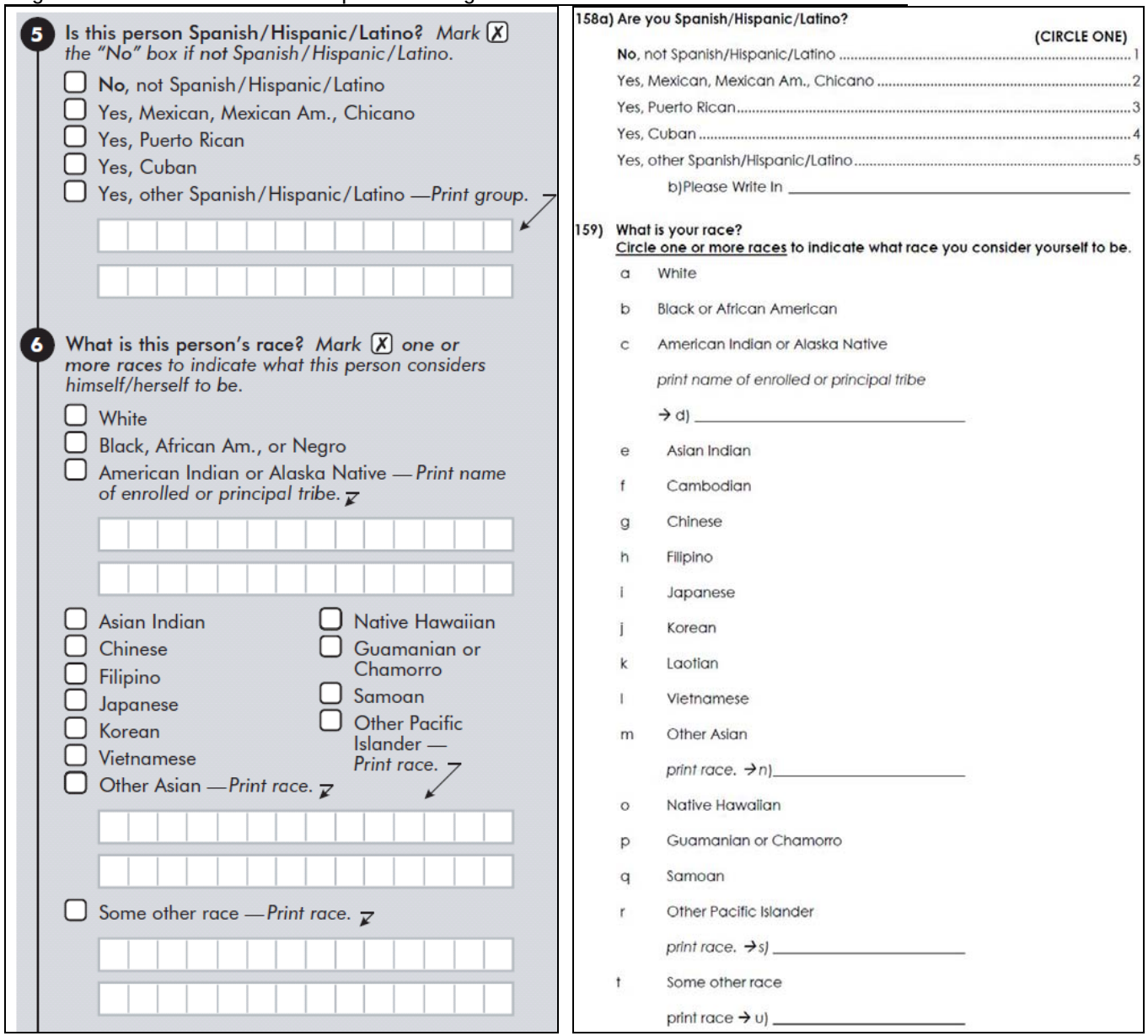
IB. Race/Ethnicity Measures in the Senior Survey

Race and Hispanic Origin

Figure 1 lists the race and Hispanic origin questions used in Census 2000 and the UW-BHS senior survey.[2] The wording, ordering, and categorization of the BHS questions are nearly identical to their Census counterparts. Both ask about Hispanic origin before race (per OMB 1997 guidelines), and both include a combination of listed categories and spaces for optional write-ins. Additionally, both sets of questions provide instructions for multiple race responses, and both include a residual category ("some other race") for respondents who fail to identify with any of the listed choices. The sole difference between Census and UW-BHS measures is the listing of additional Asian origin groups (Laotian & Cambodian) in the latter (reflecting the demographic diversity of the Pacific Northwest). Also, since the UW-BHS was administered to respondents directly (rather than to a household proxy respondent, ala Census), the question wording is written in the second person.

[2] Changes for Census 2010 and the American Community Survey are limited to the addition of a Cambodian category under Race and the omission of the word "Negro" from the list of descriptors for the "Black, African Am." category.

Figure 1-A1: Race and Hispanic origin in Census 2000 and the UW-BHS



5 **Is this person Spanish/Hispanic/Latino?** *Mark ☒ the "No" box if not Spanish/Hispanic/Latino.*

- ☐ **No**, not Spanish/Hispanic/Latino
- ☐ Yes, Mexican, Mexican Am., Chicano
- ☐ Yes, Puerto Rican
- ☐ Yes, Cuban
- ☐ Yes, other Spanish/Hispanic/Latino —*Print group.*

6 **What is this person's race?** *Mark ☒ one or more races to indicate what this person considers himself/herself to be.*

- ☐ White
- ☐ Black, African Am., or Negro
- ☐ American Indian or Alaska Native — *Print name of enrolled or principal tribe.*
- ☐ Asian Indian
- ☐ Chinese
- ☐ Filipino
- ☐ Japanese
- ☐ Korean
- ☐ Vietnamese
- ☐ Other Asian —*Print race.*
- ☐ Native Hawaiian
- ☐ Guamanian or Chamorro
- ☐ Samoan
- ☐ Other Pacific Islander — *Print race.*
- ☐ Some other race —*Print race.*

158a) Are you Spanish/Hispanic/Latino? (CIRCLE ONE)

No, not Spanish/Hispanic/Latino 1
Yes, Mexican, Mexican Am., Chicano 2
Yes, Puerto Rican 3
Yes, Cuban 4
Yes, other Spanish/Hispanic/Latino 5

b) Please Write In _____

159) What is your race?
Circle one or more races to indicate what race you consider yourself to be.

a White
b Black or African American
c American Indian or Alaska Native
print name of enrolled or principal tribe
→ d) _____
e Asian Indian
f Cambodian
g Chinese
h Filipino
i Japanese
j Korean
k Laotian
l Vietnamese
m Other Asian
print race. →n)_____
o Native Hawaiian
p Guamanian or Chamorro
q Samoan
r Other Pacific Islander
print race. →s) _____
t Some other race
print race → u) _____

Ancestry

The UW-BHS also includes an open-ended ancestry question. As shown in Figure 3, the question is worded identically (but in second person) to the one used in Census 2000,[3] though the UW-BHS lists a few additional examples.

Figure 1A-2: UW-BHS Question on Ancestry

160) What is your ancestry or ethnic origin?

_____

(For example: Italian, Jamaican, Russian, African Am., Cambodian, Cape Verdean, Norwegian, Cuban, Puerto Rican, Amer. Indian, Dominican, French Canadian, Haitian, Korean, Lebanese, Polish, Nigerian, Mexican, Taiwanese, Ukrainian, and so on.)

Mother's and Father's Ancestry and Birthplace

In addition to querying each respondent's ancestry, the UW-BHS asks respondents to report the ancestry or ethnic origin of their birth father and mother, respectively. Each of these questions is in turn followed with a query on parental birthplace. All four questions are shown in Figure 3 below.

Figure 1A-3: Parental Ancestry and birthplace in the UW-BHS

166) What is your biological/ birth mother's ancestry or ethnic origin?

_____

(For example: Italian, Jamaican, Russian, African Am., Cambodian, Cape Verdean, Norwegian, Cuban, Puerto Rican, Amer. Indian, Dominican, French Canadian, Haitian, Korean, Lebanese, Polish, Nigerian, Mexican, Taiwanese, Ukrainian, and so on.)

167) Where was she born?

_____

State or Country

169) What is your biological/ birth father's ancestry or ethnic origin?

_____

(For example: Italian, Jamaican, Russian, African Am., Cambodian, Cape Verdean, Norwegian, Cuban, Puerto Rican, Amer. Indian, Dominican, French Canadian, Haitian, Korean, Lebanese, Polish, Nigerian, Mexican, Taiwanese, Ukrainian, and so on.)

170) Where was he born?

_____

State or Country

[3] The Census 2000 long form asks "What is this person's ancestry or ethnic origin?"

Primary Identity

While the Census and most social surveys now permit the reporting of multiple races, the task of reducing multiple responses to single race categories for statutory purposes (and backwards compatibility) is rarely addressed in the questionnaire design. A few studies do attempt to resolve multiple race reports directly, typically by probing respondents who mark one or more races to choose the single race which "best" represents them. By ignoring Hispanicity, however, these "best race" questions cannot resolve ambiguity stemming from mixed Hispanic and non-Hispanic heritage (e.g. Mexican and European). Thus, while the "best race" approach provides a straightforward solution to resolving some types of mixed identities, it cannot resolve all of them, since it adheres to the problematic bifurcation of race and Hispanicity specified by the OMB standard.

The UW-BHS addresses this limitation with a broad, open ended question about primary identity, shown in Figure 4. Rather than probing only respondents who select two or more races, the primary identity question is asked of *all* respondents and comes after the inquiries on race, Hispanic origin, and ancestry. This approach allows analysts to resolve any type of mixed identities, not just those limited to the official "races" in the OMB standard.

Figure 4: Primary Race/ethnicity in the UW-BHS

161) **Considering all the ethnic and racial categories, what is your primary ethnic and/or racial identity?**

_____

Reflected Identity

While these expanded measures improve upon most survey research instruments, the race/ethnicity items described thus far all reflect the point of view of the respondent. To the extent that individuals may view themselves differently than they are viewed by others, it may be important to gauge how respondents' race/ethnic identities are perceived by observers. An easy (though perhaps unreliable) way to obtain such information is to ask respondents directly. Thus, the UW-BHS includes an open ended question that asks respondents to report how they are viewed by others, as shown in Figure 5.

Figure 5: Reflected Race/ethnicity in the UW-BHS

162) **What ethnic and/or racial category do others put you in?**

_____

IC. Other Sources of Race/Ethnicity in the UW-BHS

Administrative Records

In addition to the self-reported measures on the senior survey, the UW-BHS contains several supplemental sources of race/ethnicity data. The first is a set of

administrative records collected from respondents' schools. Unlike the more detailed measures in the senior survey, these records sort respondents into just five mutually exclusive race/ethnic categories: white, black, American Indian, Asian/Pacific Islander, and Hispanic/Latino.

While the exact source of the racial identities contained in these records is unknown, they were likely collected early at an early stage in students' educational tenure in Washington state. At minimum, the records seem to predate the 1997 revisions to the OMB standard, which specified major changes to the original (1977) version by splitting NHOPIs from Asians and allowing for multiple race responses. Neither of these changes are incorporated in the school records, and the inclusion of "Hispanic" as a coequal race/ethnic category (rather than a separate question) suggest that the data are probably not self-reported, per OMB guidelines governing self- vs. observer- identification of race & Hispanic Origin. In all likelihood, these records were created by a parent or school administrator when respondents first entered the school system.

External measures Supplement

The third distinct source of race/ethnicity data in the UW-BHS is derived from respondents' yearbook photos. These data were gathered as part of a supplemental data collection effort led by BHS researchers, the goal of which was to obtain reliable measures of respondent characteristics from independent observers. These measures were obtained using a web-based questionnaire that recorded three independent observations of race/ethnicity, attractiveness, and body type for each UW-BHS respondent. These third person appraisals (which we also refer to as "ratings") were supplied by a probability sample of current UW undergraduates (N=570), who were asked to rate 25 to 50 photos each.

Figure 6: Observed Race in the UW-BHS External measures Supplement

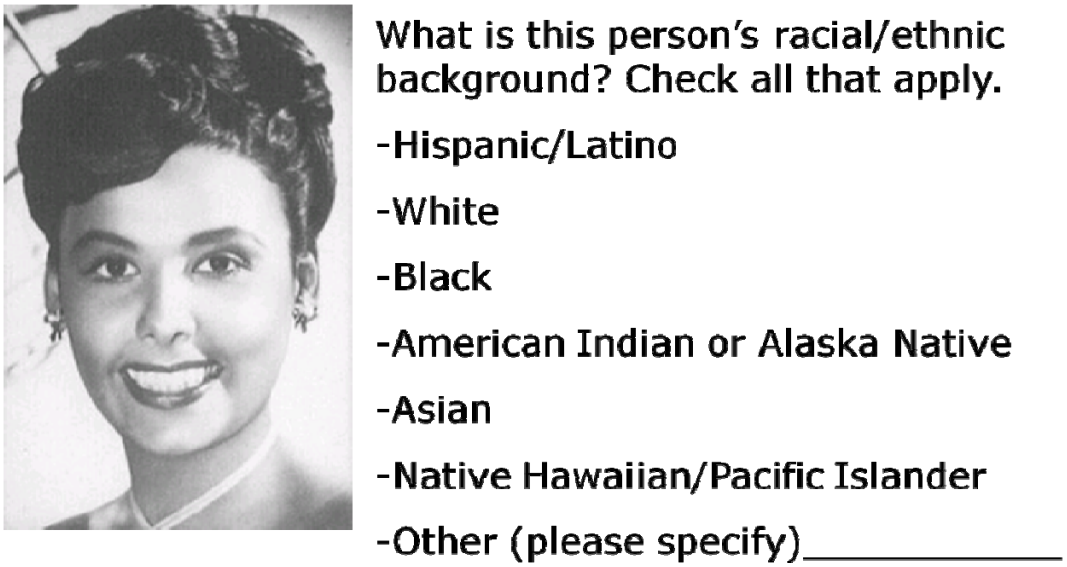


Figure 6 shows the combined race/ethnicity question used in the external measures supplement (EMS). The questionnaire wording and categorization incorporates elements of both the senior survey and administrative records. As in the latter, Hispanic origin is listed coequal to the other racial categories. This combined

format adheres to OMB standards for observer-based identification, and it spares observers the awkward task of having to assign race and Hispanic origin *separately* to each respondent.

Like the senior survey, the EMS distinguishes between Asians and NHOPIs and provides provisions for multiple responses as well as a "primary" race. In the initial query, observers are instructed to "check all that apply" when assigning race/ethnicity to the pictured respondent. Raters who select multiple categories are then prompted to identify the single category they would choose if restricted to choosing just one. This follow-up question provides the external race equivalent of the "primary identity" question (see Figure 4) from the senior survey.

The decision to gather external observations of race/ethnicity stems in part from the uncertainty surrounding the senior survey question on reflected race/ethnicity. While the latter should, in theory, measure how respondents are racially identified by others, the nature of the subject matter and the method of data collection introduce a potential bias. Not only would respondents have to know the racial/ethnic categories in which others place them, but they would also have to be willing to report those categories honestly, even if that meant acknowledging differences between how they define themselves and how they are seen by others.

The final source of race/ethnicity data in the UW-BHS is from the parental questionnaire. XXXX Due to low response rates, these data are only available for certain years (2000, 2002, and 2003) and are missing for a large share of respondents.

In all, the UW-BHS contains 9 unique measures of respondent's race/ethnic identity, excluding the parental questionnaire and race-informative queries about parental birthplace and language usage. These measures span two perspectives (first and third person) and three data sources, allowing researchers to assess race/ethnic identities, and the intersections therein, at an unprecedented level of detail. In the next section, we report basic descriptive statistics from each measure of race/ethnicity and discuss challenges in coding and interpreting these data.

II. Raw descriptives and challenges in race/ethnic measurement

As shown above, the senior survey measures race/ethnicity using either fully open ended questions or a combination of listed categories and an optional write-in space. Respondents are instructed to choose from the groups listed under a particular OMB category or supply a write-in if they identify with some other group in that category (e.g. "other Asian" or "other Hispanic"). Since the two response options measure different types of information, marked categories and write-ins are measured in separate variables. Variable names combine the number and letter from the questions in the senior survey (see section I). For example, variable **s159b** denotes those whose identify as black or African American.

IIA. Hispanic Origin: Question 158 in the Senior Survey

For Hispanic origin, all responses are coded in two categorical variables: One that measures the group each respondent circled (**s158a**), and another that records the optional write-ins they supplied (**s158b**) under "Other Spanish/Hispanic/Latino." Table 1a shows the numeric codes, code-labels, and frequency counts for each response to the five listed categories, while Table 1b lists some of the write-ins supplied in the optional space.

Most UW-BHS respondents provide a straightforward, easily coded answer to the Hispanic origin question. 84% indicate that they are not Hispanic, while another six percent identify one of the three named Hispanic categories (Mexican, PR, and Cuban). All told, 90% of respondents mark a listed category without supplying a write-in:

Table 1a: Tabulation of s158a—Hispanic Origin Categories

			Freq.	Percent
Valid	1	not spanish/hisp/latino	8128	84.16
	2	yes, mexican, mex am, chicano	413	4.28
	2.5		1	0.01
	3	yes, puerto rican	125	1.29
	3.5		1	0.01
	4	yes, cuban	25	0.26
	5	yes, other spanish/hisp/latino	336	3.48
	Total		9029	93.49
Missing	.		629	6.51
Total			9658	100.00

The remaining 10% of the sample includes those who either a) skipped the question, b) marked “Other Spanish…” and/or c) supplied multiple responses. Each of these outcomes presents unique and recurring challenges in coding Hispanic Origin and other race/ethnicity measures in the UW-BHS. Collectively, we refer to these sources of potential measurement error as the 3M’s: *Missing*, *multiple*, and *mistaken* responses.

Missing Responses

The first source of measurement error, missing responses, is an endemic problem in social research. While the 6.5% of respondents who skipped the Hispanic Origin question is in keeping with rates of item non-response to race/ethnicity queries in social surveys, it bears noting that more respondents skipped the question than identified as Mexican, Puerto Rican, and Cuban combined. Fortunately, with multiple sources of race/ethnicity data, the BHS is well suited to dealing with item non-response. By borrowing information from other race/ethnicity measures, the number cases with missing race/ethnic identities can be reduced, though this approach requires that we look at reported race/ethnic identities as a whole, rather than on a measure-by-measure basis. See the section on coding methodology for an overview.

Multiple Responses

The second challenge, multiple responses, is largely mitigated by the wording and format of the Census Hispanic Origin question, on which the UW-BHS was modeled. Since the question does not provide explicit provisions for multiple responses, we cannot identify Hispanics who circled multiple, non-adjacent categories (e.g. Mexican and Cuban), nor can we readily identify folks with mixed Hispanic and Non-Hispanic heritage. That said, a couple respondents drew a circle around adjacent categories (codes 2.5 and 3.5 in Table 1a), six respondents circled at least one *named* category and supplied an “Other Spanish…” write-in (see Table 1c below), while an even larger number supplied write-in responses that clearly indicate multiple Hispanic origins (such as "Dominican and Mexican” or "Half Mexican half Puerto Rican"). Still others supplied responses that indicate mixed ethnic backgrounds, such as “Mexican/White.” We present strategies for dealing with multiple responses to this and other race/ethnicity measures in later sections.

Mistaken Responses

The third source of error is mistaken responses, deduction of which relies on the interpretation of respondents' written identities. With an n=336, the number of Hispanics who circled "Other Spanish…" is unusually large, second only to Mexicans (n=413) in size. As shown in Figure 1, this catch-all response includes a write-in space with instructions for respondents to print their Hispanic Origin. Like similar queries for Asians and NHOPIS, the goal of the "other Spanish…" write-in is to provide a space for Latinos whose origin groups are not listed on the survey (i.e., those who are not Mexican, PR, or Cuban) to identify their specific national origins. In theory, one would therefore expect to find write-ins such as Columbian, Dominican, etc.

In practice, a variety of respondents use this space to identify themselves, and not all identities supplied are of detailed Hispanic origin groups. Table 1b presents the 20 most popular write-ins on the Hispanic Origin question, arrayed in descending order of frequency (see Appendix for unabridged version).

Table 1b: Tabulation of s158b—"Other Spanish" Write-in Responses

		Freq.	Percent
Valid	SPANISH	29	0.30
	HISPANIC	26	0.27
	PANAMANIAN	19	0.20
	PERUVIAN	15	0.16
	WHITE	12	0.12
	COLOMBIAN	9	0.09
	AFRICAN AMERICAN	8	0.08
	VENEZUELAN	8	0.08
	CHILEAN	7	0.07
	BLACK	5	0.05
	HONDURAN	5	0.05
	PORTUGUESE	5	0.05
	DOMINICAN	4	0.04
	ECUADORIAN	4	0.04
	FILIPINO	4	0.04
	ITALIAN	4	0.04
	SPANIARD	4	0.04
	UKRAINIAN	4	0.04
	AMERICAN	3	0.03
	ASIAN	3	0.03
	:	:	:
	:	:	:
	:	:	:
	Total	357	3.70
Missing		9301	96.30
Total		9658	100.00

Of the 20 most frequent write-ins, only 10 are specific Hispanic origin groups such as Panamanian or Peruvian (nine if the Portuguese are excluded, as OMB does not count Portugal or Brazil as Spanish origins). Close inspection of the write-ins reveals a multitude of unexpected results, including vague write-ins, syntax errors, redundancies, misplaced or uncodable identities, and multiethnic

identities. Some of these are mere oversights or clerical errors that can be ignored or easily fixed, while others are more difficult to resolve.

Among the former we include panethnic write-ins, syntax errors, and redundancies. First, the two most frequent write-ins reported, “Spanish” and “Hispanic,” offer no detail about respondents’ national origins, suggesting that a plurality of write-in respondents either ignored instructions to print their specific origin or simply choose to identify in vague, panethnic terms. Second, as is common with open ended questions, respondents often employ distinctive syntax (and spelling errors) when writing their identities, thereby inflating the number of unique identities reported (some of these may have been data entry errors as well). For example, although “Panama, Panamanian, and Panamainian” likely refer to the same origin, the raw values of each entry are unique in the datafile. Third, a handful of respondents supplied write-ins that replicate listed categories (i.e. wrote “Mexican American” or “Puerto Rican”), the net effect of which deflates counts of these listed categories in the raw variables. These reporting mistakes, though tedious to resolve, can be rectified with minimal assumptions and without assigning respondents to a different OMB group.

Other write-in errors are more serious and require difficult choices to resolve. Among them we include write-ins that belie *non-Hispanic* origins as well as those that are racially ambiguous. First, as shown in Table 1b, dozens of respondents wrote-in entries such as “white, black, or Asian” in the “Other Spanish” space. While these respondents would, strictly speaking, be counted as Hispanic for identifying under “Other Spanish,” it is unclear whether they are identifying *as Hispanic* or simply (mis)placing their non-Hispanic identity *in the location intended for Hispanic write-ins*. Indeed, the distinction between the content and location of identities is an enduring theme throughout this memo.

Second, while some write-ins suggest identities that are likely misplaced, others contain only ambiguous information. Entries such as “A lot of stuff,” “Not sure,” and “Adopted Hispanic” cannot be easily assigned to one of the Hispanic categories, yet it’s not clear that those who supply these responses should be presumed non-Hispanic either. The first two responses may reflect multiple Hispanic ethnicities or Hispanic origins that not fully known, while the adopted respondent may be describing his birth- or adoptive parents.

These problematic responses cannot be resolved in isolation from other sections of the question, or even other questions on the questionnaire. Thus, we propose a more holistic view of race/ethnicity, one that treats the entire survey as the measurement instrument. Not only can we then utilize multiple measures to help fill in the blanks on missing items, but we can resolve problematic responses to individual items by looking at the content of other ones. One respondent, for instance, wrote “1/16th” on the line for “Other Spanish.” Taken on its own, this information is useless. Cross-referencing this write-in with the responses to **s158a**, however, shows that the respondent also circled Mexican. Looking at the two items separately, we would be forced to treat the respondent as Mexican or designate an error code to his/her write-in. But looking at both simultaneously allows us to correctly identify the respondent as being of predominantly non-Hispanic but partial Mexican ancestry, a suspicion confirmed by subsequent write-ins on the ancestry and parental ancestry queries, which show the respondent to be of British, German, European, and Mexican descent.

Indeed, a closer look at how respondents jointly utilize the listed categories (s158a) and the write-in section (s158b) of the Hispanic origin question reiterates the contention that a number of Hispanic identity claims are probably made in error. Astute readers will note the gap between the total number of cases who

supplied a write-in (357) in Table 1b and the number that circled “Other Spanish” (336) in Table 1a. Since respondents may supply a write-in without circling “Other Spanish” and vice versa, the sums of these two items need not correspond. In this case, write-ins outnumber circled responses, suggesting that “Other Hispanics” are not the only ones using this write-in space. Table 1c illustrates the overlap between the two items in greater detail, using a simple cross-tabulation of **s158a** by a dichotomous indicator of having *any* write-in under “Other Spanish.”

Table 1c: Cross tabulation of s158a and s158b

	Supplied WI on s158b: Other Spanish…	
	No	Yes	Total
not spanish/hisp/lati	8,118	10	8,128
yes, mexican, mex am,	409	4	413
2.5	1	0	1
yes, puerto rican	124	1	125
3.5	0	1	1
yes, cuban	25	0	25
yes, other spanish/hi	21	315	336
.	604	26	630
Total	9,302	357	9,659

Results show that the overlap between circled categories and write-ins is sometimes unintuitive. 26 respondents supply a write-in without circling any category (including “other Spanish”), and another 10 supply a write-in after circling *non-Hispanic*. In addition, 21 respondents circle “Other Spanish” but provide no write-in at all.

All told, 937 respondents supplied something that could be construed as a Hispanic identity in question s158 (either by circling one of the “Yes, XXX” categories OR supplying a write-in). A large portion (378, or 40%) of that total checked “Other Spanish” or supplied a write-in underneath it, and it’s clear from the examples shown that not all who did so used the space as intended, or even in a manner that provides a clear measure of one’s Hispanic origin, if any. A number of responses are vague or ambiguous, and others offer contradictory information about whether one is Hispanic or non-Hispanic.

Resolving errors due to missing or potentially mistaken responses requires an in-depth examination of write-in content and consideration of students’ joint responses to the multiple race/ethnicity items on the UW-BHS. Indeed, w*ithout* incorporating this information, there is little hope of obtaining accurate counts of Hispanics and non-Hispanics alike. It is unclear, for instance, how to classify a student who circles “other Hispanic” only to write in “Irish” directly beneath it. Could this respondent be of mixed Irish and Hispanic origin? Or might he/she simply be misinterpreting the write-in space under “Other Spanish” as a place to self-identify *any* origins, including non-Hispanic ones? Certainly the presence of non-Hispanic write-ins in the “Other Spanish” space, some of which even follow a circled response of “No, Not Hispanic”, lends credence to this possibility.[4] In later sections, we explore several options for coding and interpreting these complex data.

[4] Some write-ins were likely provided by non-Hispanics who mistook s158 for the Race question. R's wrote "Italian", "Bosnian", or "I'm black thank you very much!" under “Other Hispanic.”

IIB. Race: Question 159 in the senior survey

A mere glance at Figure 2 portends the difficulties in coding the race question. As with Hispanic Origin, respondents are given a list of categories and space to supply optional write-ins, but the number of categories and write-in spaces are far more numerous. Further compounding matters is the inclusion of multiple national origin groups for Asians and NHOPIs, as well as respondent instructions to "mark all that apply." This format encourages multiple responses both within and between the five OMB racial categories.

There are 16 OMB-reducible categories than can be circled on the race question. Three OMB groups also include write-in spaces to print a detailed tribe (for AIANS) or specific national origin (For Asians and NHOPIs whose groups are not listed). Finally, there is a residual checkbox and write-in space entitled "some other race" (SOR) for those who fail to identify using any of the listed OMB categories or associated write-in spaces. All told, there are 21 distinct variables for the original race responses in the UW-BHS: 17 Dichotomous indicators for listed groups (coded 1 if the group was circled, 0 if it was not), and four string variables for the write-in sections. Per convention, variable names correspond to the questionnaire number and letter from the senior survey.

Table 2a: Circled Race Categories in the UW-BHS

			Percent	
Variable	**Racial Category**	**Freq.**	**Of Races Circled**	**Of UW-BHS Sample**
s159a	white	6091	55.4%	63.1%
s159b	black, african am., or negro	1359	12.4%	14.1%
s159c	american indian or alaska native	496	4.5%	5.1%
	ASIAN	2031	18.5%	21.0%
s159e	asian indian	51	0.5%	0.5%
s159f	cambodian	286	2.6%	3.0%
s159g	chinese	217	2.0%	2.2%
s159h	filipino	359	3.3%	3.7%
s159i	japanese	172	1.6%	1.8%
s159j	korean	440	4.0%	4.6%
s159k	laotian	47	0.4%	0.5%
s159l	vietnamese	318	2.9%	3.3%
s159m	other asian	141	1.3%	1.5%
	NHOPI	315	2.9%	3.3%
s159o	native hawaiian	87	0.8%	0.9%
s159p	guamanian or chamorro	44	0.4%	0.5%
s159q	samoan	121	1.1%	1.3%
s159r	other pacific islander	63	0.6%	0.7%
s159t	some other race	693	6.3%	7.2%
	Sum of Races Circled	**10985**	**100%**	**114%**
	UW-BHS Sample Size	**9659**		**100%**

Multiple Responses

Table 2a contains raw frequency counts for each of the 17 race indicator variables. The challenge presented from the instruction to “mark all that apply” is evident even before accounting for write-ins, since the number of circled categories (10,985) exceeds the number of UW-BHS respondents (9659) by more than 1300 cases. While 63% (6091) of respondents circled white, for instance, this represents only 55% of the total number of races circled. Still, most of the circled identities are major OMB categories like white, black, and American Indian, which alone account for more than 72% of all responses tallied. Another 17.2% and 2.3% circle one of the named Asian or NHOPI groups (like Chinese and Samoan), respectively, while 2% circle one of the residual categories for “Other” Asians/NHOPIs.

Table 2b: Number of Categories Circled on s159a

		Freq.	Percent	Valid	Cum.
Valid	0	573	5.93	5.93	5.93
	1	7654	79.24	79.24	85.17
	2	1104	11.43	11.43	96.60
	3	253	2.62	2.62	99.22
	4	56	0.58	0.58	99.80
	5	10	0.10	0.10	99.91
	6	3	0.03	0.03	99.94
	8	1	0.01	0.01	99.95
	9	2	0.02	0.02	99.97
	12	1	0.01	0.01	99.98
	17	2	0.02	0.02	100.00
	Total	9659	100.00	100.00	

Table 2b arrays the number of categories chosen by the frequency of respondents who chose them. The cross-product of the two sums to the 10,985 total responses. As shown in the table, only 79% of respondents circled a single category, while 11.4% and 2.6% marked two and three categories, respectively. About 0.75% of the sample marked between 3-12 categories, while two respondents circled all 17.

While it is clear that many respondents marked multiple categories, these simple tabulations cannot be used to determine the number of multiracial identities, per se. First, these counts exclude write-ins, which may contain additional, OMB-reducible identities (Malaysian, French, Nigerian, etc.). Second, these counts treat “some other race” coequal to the OMB categories. A respondent who circles white and SOR but writes “Irish” is only identifying one racial category, not two. Third, the listing of sub-OMB groups (e.g. Korean, Chinese) makes it impossible to distinguish multiracial responses (e.g. Asian and black) from multi-ethnic ones (e.g. Chinese and Japanese). Each of these factors clouds efforts to know precisely how many respondents identify which each OMB group, much less whether they identify with that group alone or in combination with other OMB groups. Only by looking at the combinations of identities circled, and coding those which are written in, can these issues be fully resolved.

Missing Responses

As shown in Table 2b, almost 6% of the sample failed to circle even one category. Moreover, Table 2a shows another 6.3% of the sample identifies as “some other race.” The correspondence between these figures on rates of non-response to question 159 cannot be determined without looking at the content of the write-ins, as well as the combination of identities reported. Some respondents who fail to circle a category may supply an OMB-codable write-in (Irish => white, e.g.). On the

other hand, some of those whose identify under SOR may not supply a response that can be coded to an OMB category ("Elf" => ????), and could therefore be coded as missing.

Mistaken Responses

Errant responses can also be ascertained only by looking at write-ins and combinations. There is no way to invalidate the identities of respondents who circled white, black, or one of the named Asian or NHOPI categories. Even if respondents circled AIAN or one of the "other" Asian/NHOPI categories, but failed to supply a write-in (as instructed), their responses must be treated "as is," since the absence of specific information cannot invalidate claim of membership within the broader OMB group. We can still list these respondents as generic Asian, NHOPI, or AIAN, and indeed, the might very well identify as such.

But what should be done with respondents whose Indian tribe is "Polish," or whose Asian ancestry is "French?" The write-in spaces were designed to provide an opportunity for respondents to identify unlisted Asian or NHOPI groups or list their tribe if they are AIAN. In practice, respondents' answers often deviate from these intentions, much as they do with the "Other Spanish" write-in.

Table 2c: Sample Write-in errors

			Types of Write-in Errors				
OMB Group	**Write-in Variable**	**Freq.**	**Syntax**	**Vague Identity**	**Multiple Identity**	**Ambiguous Identity**	**Misplaced Identity**
AIAN	s159d	327	CHALKTAH	NATIVE AMERICAN	BLACK, WHITE, INDIAN	LIKE I KNOW	POLISH AND GERMAN
Asian	s159n	140	LAOTIAN (LAO)	ASIAN AMERICAN	KOREAN, FRENCH	DK ADOPTED	FRENCH, IRISH
NHOPI	s159s	58	FIJIAN SEVENTY FIVE PERCENT	POLYNESIAN	WHITE & HAWAIIAN	MIXED	FILIPINO

Table 2c provides counts of write-ins and examples of write-in errors for each of the three OMB categories with an optional write-in space (see Appendix for unabridged version). 327 respondents supplied an AIAN write-in, versues just 140 Asians and 58 NHOPIs. The larger number of AIAN write-ins owes to the fact that that no tribes were listed as categories, while several Asian and NHOPI origin groups were.

Since write-ins and circled responses are coded separately, the totals will not correspond perfectly. As with Hispanic origin, the numbers who circle the write-in category differ from the numbers who write something in the space underneath it. Perhaps the most glaring discrepancy on Q. 159 is the gap between the 496 respondents who circled AIAN (Table 2a) and the 327 who then followed the instructions to print their tribe. Clearly, not all who identify as AIAN can or

will name a specific tribal ancestry. Even the 327 respondents who write something in inflates the number of AIANs with a specific tribal identity, since some of the write-ins are vague, ambiguous, or misplaced, as shown in Table 2c and the unabridged table in the Appendix.

Indeed, while most respondents use the race write-ins as intended, others commit similar errors to those observed in the “Other Spanish” write-in. In addition to vague and misspelled identities, several write-ins are ambiguous (don’t know, mixed, etc.) or misplaced (Polish listed as an Indian tribe). While clerical errors (typos and strange syntax) can be fixed with a simple cleaning, the more substantive errors cannot be resolved without looking beyond the individual write-ins to the full array of race/ethnic identities supplied on question 159 (and beyond). In so doing, it may become necessary to reassign respondents from one OMB group to another, particularly if we opt to privilege *what* the respondent writes (Japanese, e.g.) over *where* they write it (in the space meant for Indian tribes, e.g.), again highlighting the distinction between the content and location of identities.

Despite the inclusion of multiple listed categories and follow-up write-in spaces, 693 respondents (7% of the sample) circled “some other race.” An even larger number (702) wrote something in the SOR write-in space. Table 2d details the 20 most frequent responses (unabridged version in the Appendix).

Table 2d: Truncated Tabulation of s159u -- Other race write-ins

		Freq.	Percent	Valid	Cum.
Valid	HISPANIC	71	0.74	10.11	10.11
	MEXICAN	67	0.69	9.54	19.66
	ITALIAN	35	0.36	4.99	24.64
	PUERTO RICAN	25	0.26	3.56	28.21
	GERMAN	18	0.19	2.56	30.77
	IRISH	17	0.18	2.42	33.19
	RUSSIAN	17	0.18	2.42	35.61
	LATINO	15	0.16	2.14	37.75
	HUMAN	13	0.13	1.85	39.60
	MEXICAN AMERICAN	12	0.12	1.71	41.31
	AMERICAN	9	0.09	1.28	42.59
	GREEK	8	0.08	1.14	43.73
	PERSIAN	8	0.08	1.14	44.87
	UKRAINIAN	7	0.07	1.00	45.87
	FRENCH	6	0.06	0.85	46.72
	MIXED	6	0.06	0.85	47.58
	POLISH	6	0.06	0.85	48.43
	SPANISH	6	0.06	0.85	49.29
	ARAB	5	0.05	0.71	50.00
	CREOLE	5	0.05	0.71	50.71
	:	:	:	:	:
	Total	702	7.27	100.00	100.00
Missing		8957	92.73		
Total		9659	100.00		

Responses to the SOR write-in are distributed across a large number of entries. Collectively, the top 20 write-ins account for only half of the 700+ total write-ins supplied, and only 10 write-ins contain a double-digit number of respondents. Summarizing these data will require tedious coding of various categories and combinations, many of which are simply distinct variants of a common origin.

While we detail these coding procedures in later sections, a brief inspection of the SOR entries reveals several patterns. First, three of the four most frequent write-ins are Hispanic. The entries “Hispanic; Mexican; and Puerto Rican” alone account for nearly a third of all SOR write-ins, and nearly half if variants like “Mexican American” or “Latino” are added.

Second, many popular write-ins are redundant with listed categories or subsumed by major OMB groups, including white (“German, Irish, French,” etc.) black (“African”), and Asian (“Cambodian, Korean” etc.). While the decision to ignore listed categories might reflect respondents’ recent immigration experience, rejection of pan-ethnic labels, or a simple oversight, excluding these write-ins biases counts of listed categories downward.

Third, in addition to overlooking the listed categories, some respondents seem to have ignored the “circle one or more races” instruction and instead used the SOR space to identify their mixed race/ethnic roots. “Creole” and “mixed” are two popular write-ins of this type, and several respondents supplied answers such as “black and white” or “biracial” (unabridged table in the Appendix).

Finally, a number of respondents were either unwilling or unable to identify an easily codable identity. Entries of this type range from the unsure (I wish I knew, Don’t know, etc.) to the absurd (Elf, Medium rare, etc.), and some even suggest a subtle aversion to the idea of racial categorization in general (Human, Homo sapiens, etc.). One respondent even goes so far as to indict the BHS staff for including race/ethnicity queries on the survey, using the SOR space to demand: “STOP ATTACHING RACE TO HOW A PERSON THINKS, SICKOS.”

As these examples make clear, circled responses to self-reported race/ethnicity queries often fail to enumerate their target populations. Write-ins must be coded and incorporated to obtain more accurate estimates of subpopulation size. Many entries are easily coded to major OMB groups, while others beg the question of which OMB race, if any, should be assigned. Designating an OMB code to ambiguous or averse respondents would normally be impossible, but the multiple race/ethnicity measures in the senior survey and other BHS sources provide insight to the identities of even the most stubborn respondents, without having to rely on imputation or casewise deletion.

IIC. S160, s166, and s169: Ancestry and Parental Ancestry

In addition to the standard Census questions on race and Hispanic origin, the UW-BHS contains a number of open-ended race/ethnicity queries, introduced above. The following pages present raw descriptives and challenges for five of these measures: Ancestry, mother’s ancestry, father’s ancestry, primary identity, and reflected identity.

Table 3a lists the 20 most popular write-ins for the open-ended ancestry question (variable **s160**), while Tables 3b and 3c do the same for mother’s and father’s ancestry (**s166** and **s169**, respectively). At first glance, the questions appear to have worked as intended. Most of the popular entries contain detailed origin groups, only some of which are typically included on race and Hispanic Origin questions. Clearly, the use of additional race/ethnicity measures provides a far more nuanced perspective on population diversity, and open-ended queries can enumerate far more categories and combinations, while taking up less space on the survey, than their close-ended counterparts.

The trade-off is data that are far more difficult to process and interpret. Upon inspection, it becomes clear that the challenges encountered in the s158 and s159

write-ins are magnified several fold in these fully open-ended questions. For example, many ancestry write-ins provide no detail beyond the broad OMB categories (African American, White), and thus offer little improvement over standard race/ethnicity measures. Other respondents refuse to supply even OMB-level detail about their origins (or their parents'). "American" is a top 20 response for all three queries. Still others (not shown) supply an abundance of detail, listing two or more groups at a time, making comparisons between ethnic groups difficult. "German, Irish" is the 20th most popular ancestry reported, for example.

Some responses (see Appendix) are ambiguous and can be coded to two or more OMB categories (e.g., "Indian"), while others indicate respondent fatigue with the repeated queries about ethnic background in the senior survey. One disgruntled student writes: "F RACIAL IDENTITY, I AM ME, WHO CARES." Perhaps most notably, rates of non-response to the open ended items are considerably higher. Tables 3a-c show that 11%, 13%, and 16% of the sample skip the questions on ancestry, mother's ancestry, and father's ancestry, respectively (though the latter may owe, in part, to limited contact with fathers.)

Table 3a: Tabulation of s160, Ancestry or ethnic origin

		Freq.	Percent	Valid	Cum.
Valid	AFRICAN AMERICAN	431	4.46	5.01	5.01
	KOREAN	270	2.80	3.14	8.15
	GERMAN	265	2.74	3.08	11.24
	MEXICAN	213	2.21	2.48	13.71
	VIETNAMESE	207	2.14	2.41	16.12
	CAMBODIAN	182	1.88	2.12	18.24
	IRISH	178	1.84	2.07	20.31
	FILIPINO	151	1.56	1.76	22.07
	NORWEGIAN	148	1.53	1.72	23.79
	ITALIAN	132	1.37	1.54	25.32
	WHITE	127	1.31	1.48	26.80
	AMERICAN	106	1.10	1.23	28.03
	EUROPEAN	64	0.66	0.74	28.78
	CHINESE	58	0.60	0.67	29.45
	RUSSIAN	57	0.59	0.66	30.12
	AFRICAN	54	0.56	0.63	30.74
	SAMOAN	53	0.55	0.62	31.36
	ENGLISH	50	0.52	0.58	31.94
	UKRAINIAN	47	0.49	0.55	32.49
	GERMAN, IRISH	46	0.48	0.54	33.02
	:	:	:	:	:
	Total	8597	89.01	100.00	100.00
Missing		1062	10.99		
Total		9659	100.00		

Table 3b: Tabulation of s166, Mother's Ancestry or ethnic origin

		Freq.	Percent	Valid	Cum.
Valid	GERMAN	498	5.16	5.97	5.97
	AFRICAN AMERICAN	391	4.05	4.68	10.65
	WHITE	389	4.03	4.66	15.31
	KOREAN	386	4.00	4.62	19.94
	IRISH	315	3.26	3.77	23.71
	NORWEGIAN	235	2.43	2.82	26.52
	VIETNAMESE	227	2.35	2.72	29.24
	MEXICAN	212	2.19	2.54	31.78

	CAMBODIAN	191	1.98	2.29	34.07
	FILIPINO	191	1.98	2.29	36.36
	ITALIAN	168	1.74	2.01	38.37
	ENGLISH	133	1.38	1.59	39.97
	BLACK	127	1.31	1.52	41.49
	AMERICAN	112	1.16	1.34	42.83
	SWEDISH	79	0.82	0.95	43.78
	FRENCH	78	0.81	0.93	44.71
	EUROPEAN	76	0.79	0.91	45.62
	RUSSIAN	67	0.69	0.80	46.42
	CHINESE	66	0.68	0.79	47.21
	JAPANESE	66	0.68	0.79	48.01
	:	:	:	:	:
	Total	8347	86.42	100.00	
Missing		1312	13.58		
Total		9659	100.00		

Table 3c: Tabulation of s169, Father’s Ancestry or ethnic origin

		Freq.	Percent	Valid	Cum.
Valid	AFRICAN AMERICAN	530	5.49	6.55	6.55
	GERMAN	516	5.34	6.38	12.93
	WHITE	352	3.64	4.35	17.28
	IRISH	273	2.83	3.37	20.65
	MEXICAN	245	2.54	3.03	23.68
	KOREAN	241	2.50	2.98	26.66
	VIETNAMESE	209	2.16	2.58	29.24
	ITALIAN	196	2.03	2.42	31.66
	BLACK	187	1.94	2.31	33.97
	NORWEGIAN	187	1.94	2.31	36.28
	CAMBODIAN	180	1.86	2.22	38.51
	FILIPINO	170	1.76	2.10	40.61
	AMERICAN	143	1.48	1.77	42.38
	ENGLISH	117	1.21	1.45	43.82
	SCOTTISH	85	0.88	1.05	44.87
	CHINESE	74	0.77	0.91	45.79
	FRENCH	68	0.70	0.84	46.63
	RUSSIAN	66	0.68	0.82	47.44
	SAMOAN	64	0.66	0.79	48.23
	POLISH	63	0.65	0.78	49.01
	:	:	:	:	:
	Total	8092	83.78	100.00	
Missing		1567	16.22		
Total		9659	100.00		

Each of these types of measurement error complicates efforts to code the open-ended responses into a usable number of discrete categories. While the level of nuance in respondents’ backgrounds is strength of these data, a simple numeric encoding of each entry, as is, would be worthless. Not only would categories that differ only in spelling or syntax (e.g. Italy vs. Italian) receive different codes, but the sheer number of distinct codes would make the variables unusable. The small cumulative percentages from tables 3a-3c foretell the problem. As shown in Table 3a, the top 20 ancestries collectively account for just 33% of the UW-BHS sample, meaning the balance of respondents are disbursed over a large number of small, distinct categories (race/ethnic combinations, unique spelling and syntax, etc.).

Table 3d. Unique Responses to Open-ended Race/ethnicity Q’s

Variable	Description	Valid Obs.	Unique Categories	Average Obs. Per Category
s160	Ancestry	8597	4052	2.12
s166	Mother's Ancestry	8347	2412	3.46
s169	Father's Ancestry	8092	2239	3.61

Table 3d highlights the scope of the problem by listing the number and size of unique race/ethnicity "groups" and the average size of each. The results indicate that respondents identify their roots using a dizzying number of distinct combinations and wordings. For ancestry alone, respondents identify *4052 unique categories*. While eliminating typos/spelling errors and combining syntax variants (African American v. African Amer.) will reduce the number of categories, it's clear that the process of cleaning, coding, and summarizing these data presents a daunting challenge. To put the variability of these responses in perspective, simply divide the number of non-missing observations by the number of unique categories. In so doing, we learn that the average size of each write-in group is just 2-4 cases for ancestry and parental ancestry questions.

IID. Primary Identity and Reflected Identity

Table 3e shows the 20 most frequent responses to the primary identity question, variable **s161**. Results suggest the measure performs largely as advertized. The top 20 responses account for a solid majority (69%) of the sample, and all are limited to a single category. Unlike the detailed ancestry responses, more than 40% of respondents list a pan-ethnic or racial category (white, black, Asian, etc.) as their primary identity.

Still, most of the now familiar coding challenges remain. Wording variants again deflate counts of major categories (e.g. White v. Caucasian, Black v. African American), and though the number of unique entries is appreciably fewer than those observed for ancestry and parental ancestry, there all still 1296 distinct write-ins. Hundreds of these entries are ambiguous, uninformative, or otherwise difficult to code, and some respondents ignore the goal of the question and instead list *multiple* primary identities. Also, in keeping with the higher rates on non-response to the open-ended race/ethnicity questions, 13% of cases are missing.

Table 3e: Tabulation of s161--Primary ethnic/racial identity

		Freq.	Percent	Valid	Cum.
Valid	WHITE	2002	20.73	23.84	23.84
	AFRICAN AMERICAN	495	5.12	5.90	29.74
	CAUCASIAN	454	4.70	5.41	35.15
	BLACK	351	3.63	4.18	39.33
	IRISH	281	2.91	3.35	42.68
	GERMAN	276	2.86	3.29	45.96
	KOREAN	249	2.58	2.97	48.93
	AMERICAN	206	2.13	2.45	51.38
	CAMBODIAN	200	2.07	2.38	53.76
	FILIPINO	186	1.93	2.22	55.98
	MEXICAN	183	1.89	2.18	58.16
	ITALIAN	181	1.87	2.16	60.31
	VIETNAMESE	179	1.85	2.13	62.45
	NORWEGIAN	134	1.39	1.60	64.04
	ASIAN	99	1.02	1.18	65.22

	HISPANIC	82	0.85	0.98	66.20
	ENGLISH	77	0.80	0.92	67.12
	SAMOAN	60	0.62	0.71	67.83
	CHINESE	52	0.54	0.62	68.45
	SWEDISH	49	0.51	0.58	69.03
	:	:	:	:	:
	Total	8396	86.92	100.00	
Missing		1263	13.08		
Total		9659	100.00		

Importantly, two of the most frequent primary identities are “Mexican” and “Hispanic,” confirming earlier suspicions that number a number of respondents view Hispanic origin, rather than one of the five OMB “races,” as their most salient identity. This valuable information would normally be unattainable, since race and Hispanic origin are measured separately in the standard Census questions. Even surveys that include “best race” follow-ups are limited to multiple *race* responses, rather the mixture of race and Hispanic origin.

Table 3f: Tabulation of s162—-Reflected Race/ethnicity

		Freq.	Percent	Valid	Cum.
Valid	WHITE	2938	30.42	35.79	35.79
	CAUCASIAN	453	4.69	5.52	41.30
	AFRICAN AMERICAN	445	4.61	5.42	46.72
	BLACK	392	4.06	4.77	51.50
	ASIAN	331	3.43	4.03	55.53
	AMERICAN	196	2.03	2.39	57.92
	KOREAN	196	2.03	2.39	60.30
	MEXICAN	180	1.86	2.19	62.50
	IRISH	132	1.37	1.61	64.10
	VIETNAMESE	118	1.22	1.44	65.54
	CAMBODIAN	117	1.21	1.43	66.97
	GERMAN	112	1.16	1.36	68.33
	HISPANIC	105	1.09	1.28	69.61
	FILIPINO	96	0.99	1.17	70.78
	ITALIAN	93	0.96	1.13	71.91
	RUSSIAN	59	0.61	0.72	72.63
	PACIFIC ISLANDER	48	0.50	0.58	73.22
	NONE	44	0.46	0.54	73.75
	NORWEGIAN	44	0.46	0.54	74.29
	CHINESE	43	0.45	0.52	74.81
	:	:	:	:	:
	Total	8210	85.00	100.00	
Missing		1449	15.00		
Total		9659	100.00		

Table 3f details the most popular responses to the final open-ended question on the senior survey, reflected race/ethnicity (variable **s162**), which asks respondents to report how others identify them. Responses echo both the strengths and weaknesses of the primary identity question. 3/4 of respondents are clustered in the top 20 categories, and white (or Caucasian), black (or African American), Asian, and “American” collectively account for 58% of all write-ins. This makes sense since observers are less likely to know the particulars of an individual’s ancestry and are therefore more likely to classify others in simple, pan-ethnic terms. The trend toward a smaller of number of large, pan-ethnic categories shows that respondents are willing to acknowledge this simplification while reporting who others see them.

The data are far from perfect, however. Many respondents appear to have no reflected racial identity. 15% of cases are missing (2nd only to father's ancestry), and both "American" and "none" are in the top 20. In addition, the many challenges uncovered in the ancestry and primary identity questions remain. Many respondents supply multiple or ambiguous identities, and there are still 1176 unique entries reported.

Of course the biggest question surrounding the reflected identity question is whether respondents are accurately relaying external appraisals of the identities. To help resolve this question, we turn to the two supplemental sources on race/ethnicity used in the BHS, administrative records and high school yearbooks.

IIE. Administrative and External race

Table 4a details respondents' race/ethnic identities from school records. As discussed above, these records combine race and Hispanic origin (permitted under OMB guidelines when race is not self reported), and include only the broad OMB categories. Since most of these records originated prior to the 1997 revisions. Asians and Pacific Islanders are combined into a single category (API), and neither multiple responses nor SORs are recorded. In all, administrative race records are available for nearly 88% of UW-BHS respondents. The original variable is called **schlrace**.

Table 4a: Race/ethnicity from School Administrative Records

		Freq.	Percent	Valid	Cum.
Valid	1 asian	1455	15.06	17.18	17.18
	2 afric amer	1245	12.89	14.70	31.89
	3 hispanic	426	4.41	5.03	36.92
	4 natv amer	129	1.34	1.52	38.44
	5 white	5212	53.96	61.56	100.00
	Total	8467	87.66	100.00	
Missing	.	1192	12.34		
Total		9659	100.00		

The lack of write-ins and multiple responses and the inclusion of Hispanic as coequal category provide race data that are simple to interpret and analyze. Unfortunately, these simplifications also prevent easy comparison to the more complex senior survey measures. Only after collapsing the survey data can we attempt comparisons between the two sources, and these comparisons must be made with caution. We don't know who filled out these records or how the identities might have differed if race and Hispanic origin were asked separately or if multiple responses were permitted. Nor do we know if missing cases are missing at random, or if some groups are systematically underrepresented. As a result, the greatest value of the administrative records may be in their use as tie breakers for multiple responses or imputed values for missing or ambiguous self-reports.

The data on external race/ethnicity combine elements from the administrative measures as well as the senior survey. Like the former, Hispanic is included as coequal racial category, and only major OMB groups are included (see Figure 6). As in the senior survey, raters (the observers who viewed and assigned race and other characteristics to respondents' photos) were permitted to check all the major races they thought the photographed respondent might be, and raters who checked multiple categories were given a follow-up asking for the single category they would choose if limited to just one. Raters could also check SOR and supply a write-in if they

felt the pictured respondent couldn't be classified using any of the listed categories. In all, raters could assign up to seven categories for each UW-BHS respondent (white, black, AIAN, Asian, NHOPI, Hispanic, and SOR). 8441 respondents have a valid yearbook photo, so external race measures are available for roughly 88% of the senior survey sample.[5]

The use of a web-based instrument allowed us to minimize certain forms of measurement error in the EMS. All photos were rated, so there are no missing data, and errant multiples or write-ins were kept in check by the EMS design. For example, unlike the senior survey, those who checked SOR on the EMS were forced to supply a write-in, and those who supplied a write-in could only do so after marking SOR. In addition, raters who checked multiple categories were then prompted to pick from among those categories (and *only those* categories) the single identity that best reflects what they think the respondent might be. This eliminates the possibility of having multiple "primary" race/ethnic identities or a primary identity that was not included in the original combination (e.g., marking white and black but then saying "best" race is Asian, an error we could not prevent in the senior survey).

While these design choices improve the quality of the EMS data, the decision to collect multiple ratings on each photograph adds complexity. Pre-testing of the EMS instrument revealed a fair degree of inter-rater variability. Nonetheless, we determined that reliable measures of external race (those on which a majority of raters agree) could be obtained with as few as three independent ratings. Efforts to meet this target were largely successful. 8151 of the 8448 yearbook photographs were viewed by three unique raters and given a valid rating by each one. 295 photos required a fourth viewing because A) one of the raters quit the EMS before supplying a rating for that photo or B) the photo was inadvertently shown twice by the same rater. In both scenarios, a 4^{th} rating was necessary to obtain at least three independent measures. Due to the manner in which we flagged cases that had missing or duplicate ratings, we failed to obtain 3 independent ratings for a pair of particularly unlucky cases that had *two* of the three original observers quit the EMS before rating that exact picture.

All told, with 2 respondents getting 2 ratings each, 8151 respondents getting 3 ratings each, and 295 respondents getting 4 ratings each, we obtained a grand total of 25,637 valid ratings for the 8848 photographs that were observed, roughly 3.03 ratings per photo.

The total number of *races* observed is naturally higher than the total number of ratings (25,637), since observers could assign multiple races to each pictured respondent. In all, 29,455 races were assigned to the pictures viewed. Dividing this sum by the total number of ratings (25,637) shows that each in each rating, an average of 1.15 races was assigned to the respondent pictured. This per capita tally is similar to the average number of self-reported categories per person, 1.14 (sum of races tallied from Table 2a).

Table 4b: Raw tabulation of External race indicator varibles

[5] 8448 pictures were actually shown, seven of which were inadvertent (no senior survey data)

Original OMS Variable	Observed Race	Total Races Assigned (1)	Percent
picwht	White	15598	53.0%
picblk	Black	3785	12.9%
picami	AIAN	1300	4.4%
picasn	Asian	3731	12.7%
picopi	NHOPI	1491	5.1%
picsor	SOR	84	0.3%
pichsp	Hispanic	3466	11.8%
	Total	29455	100.0%

(1) Sum of all races assigned by all raters. Dramatically exceeds number of UW-BHS respondents since each respondent was rated multiple times, and each rater could assign multiple races

Table 4b presents raw tabulations for each of the seven dichotomous race/ethnic variables included in the EMS. While we save detailed comparisons of self- and –external race for later sections, one quick point emerges from the table: SOR responses are almost non-existent. This likely owes to the inclusion of Hispanics in the combined race/ethnicity question used in EMS, vindicating our decision not to ask separate questions on race and Hispanic origin. Clearly, including Hispanic as race results in a smaller number of ambiguous, difficult to code write-ins in the SOR section.

Comparisons of the distributions in Tables 4b (external race) and 2a (self reported race) should be drawn with caution, since the number of categories and the level of detail vary. The variables in Table 2a do not include responses to the Hispanic origin question and do not distinguish between multi-ethnic (within OMB) and multiracial (between OMB) responses, whereas the external race measure combines race and Hispanic origin but contains no OMB sub-groups. In addition, neither table codes SOR write-ins, so the true counts of OMB categories remain unknown for both measures.

III. Data Processing Methodology

Many of the limitations in measuring race/ethnicity are addressed by the UW-BHS study design, which includes multiple measures of identity gathered from multiple perspectives and data sources. But while these design choices are a strength of the study, the resulting data are highly complex and difficult to interpret in raw form, as shown in the many examples described above. To help users fully utilize these data, we outline a comprehensive methodology to process and interpret the multiple race/ethnicity measures in the UW-BHS.

IIIA. Overview

Multiple measures provide more opportunity for respondents to provide a race/ethnic identity, but they also provide more opportunity for measurement error. Thus, design improvements alone cannot solve all the problems of race/ethnic measurement, and complex measures such as those used in the UW-BHS create additional problems of their own.

The first challenge stems from our efforts to maintain comparability with standard Census measures of race and Hispanic origin. This requires that we adhere to an OMB taxonomy that separates race and Hispanic origin, allows multiple responses to the former but not the latter, and provides seemingly arbitrary guidelines for simplifying mixed race identities.

UW-BHS design improvements address some of these problems. We provide opportunities for respondents to treat race and Hispanicity as equals in open-ended questions on ancestry and parental ancestry, and we attempt to simplify multiple identities, including mixed race/Hispanic identities, with a question on primary race/ethnicity. But these innovations create measurement challenges of their own. The use of multiple self-administered, open-ended (or partially open-ended) questions increase fatigue and invite an “anything goes” reaction from respondents, who often choose to ignore questions and/or use them in unintended ways. As discussed above, we encounter a host of problematic responses, including vague write-ins, syntax errors, redundancies, misplaced or uncodable identities, and missing data. The sheer number of unique write-ins makes many variables unusable in raw form, and the errors and inconsistencies within and between different senior survey questions makes it difficult to compare identities from one measure to the next, or even obtain a coherent identity for each respondent.

Also, while the use of supplemental data improves upon the limited perspective of the self-reported race/ethnicity measures that predominate in social research, differences in question wording and format leave us with administrative and external race data that are not directly comparable to the senior survey measures. Both supplemental sources combine race and Hispanic origin, and neither includes sub-OMB categories. Apples-to-apples comparisons of the self- and observed- race measures will require considerable processing of the raw data from each source.

Without such processing, the richness and multidimensionality of the UW-BHS design is unlikely to be seen as a strength. Our study presents users with dozens of potential race/ethnicity measures collected at multiple time points and from multiple perspectives. These measures vary in format, quality, and complexity. Responses to individual questions are often ambiguous, and responses to different questions (or even different parts of the same question) may offer contradictory information. The combined results of these existent and emergent challenges are data that trade-off the narrow scope and theoretical naiveté of more limited race/ethnicity measures for a broader, richer set of measures that are highly (perhaps hopelessly) complex and nearly impossible to interpret in raw form.

Refining these data requires a multistep process. We begin by outlining the logic and justification of our conceptualization of race and ethnicity, which holds that the distinctions between the two concepts are less than clear. After laying out our integrated race/ethnic taxonomy, we outline procedures for coding and interpreting the multitude of responses to race/ethnicity queries in the BHS. In so doing, we make an explicit distinction between the content and location of race/ethnic identities and provide tables to illustrate the overlap and divergence between the two. Finally, we present and explain a series of constructed variables that summarize race/ethnicity at different levels of detail, from a variety of perspectives, and under several alternative assumptions.

Overlap between Race and Hispanic Origin

The distinction between race and ethnicity is, in large part, arbitrary. While the two can be viewed as conceptually distinct (race is more often associated with physical appearance, e.g.), public perceptions and legal definitions often treat

two concepts interchangeably. Many white ethnic groups were considered distinct "races" in 19th and early 20th century U.S., and many ethnic and national origin groups (Chinese, Korean, etc.) are listed as races on current census forms. Likewise, Hispanics, the sole "ethnic" group defined by the OMB standard, are often viewed as a de facto racial category (comparable to non-Hispanic whites, blacks, etc.) among Latinos and non-Latinos alike.

Indeed the very wording of the OMB classification defines both "race" and "ethnicity" in terms of geographic origin or descent. Asian or Pacific Islander is defined as "a person having origins in any of the original peoples of the Far East, Southeast Asia, the Indian subcontinent, or the Pacific Islands," while Hispanics are defined as anyone who can trace their origins to Spain (or a former Spanish colony like present day Mexico).

Evidence from the UW-BHS reiterates the similarities between race and ethnic identities. Some of the most popular write-ins supplied on the ancestry and parental ancestry questions are simple racial categories such as white and black (see Tables 3a-3c). Also, as Tables 1b and 2d amply illustrate, non-Hispanics often identify in the space reserved for "Other Spanish" on the Hispanic origin question, while Hispanics do the same in the "Other Race" section of the race question. This reciprocal spillover between race and Hispanicity inflates counts of "Other" responses on both questions, deflates counts of listed OMB categories, and results in differential rates of non-response to the two queries.

Table 5: Missing and SOR Race responses by Hispanic Origin

	Race = Missing				Race = SOR		
Group	**Yes**	**No**	**Tot.**		**Yes**	**No**	**Tot.**
Not Hispanic in s158	0	8,362	8362		406	7,956	8362
Hispanic in s158	191	746	937		306	631	937
Missing on s158	360		360			360	360
			9659				**9659**

Table 5 confirms these suspicions by cross-tabulating raw responses to s158 (Hispanic origin) and s159 (race).[6] The first column compares counts of non-response to the race question among self-identified Non-Hispanics and Hispanics, while the second compares counts of SOR responses for the same two groups. Results show that self-identified Hispanics are drastically overrepresented in each problem category. Not one self-reported non-Hispanic skips the race question, and fewer than 5% (406/8362) identify as SOR. Among self-reported Hispanics, one in five (191/937) skips the race question, and one in three (306/937) identifies as "other race." Clearly, a sizeable share of Hispanics see their "race" as Hispanic, and will either pass or say "none of the above" in response to a race question that doesn't include Hispanic origin as a category.

Race/ethnic taxonomy in the UW-BHS

In light of the conceptual and empirical blurriness of race and ethnicity, we propose an integrated taxonomy that treats the two concepts more or less

[6] The row variable is **idinhsp** (coded 0 for those who check non-Hispanic, 1 for those who circle a Hispanic category or supply a write-in under "Other Spanish," missing for those who do neither), while the column variables are **idin159** (coded 1 for those who circle a race or supply any write-in on s159, 0 for those who don't) and **idinsor** (coded 1 for those circle or write-in under SOR)

interchangeably. This hierarchical classification, shown in full in the Appendix, summarizes respondents' identities at three levels of detail, ranging from broad panethnic/racial categories to specific nationalities and tribes. This taxonomy can be applied to any combination of marked or written responses, and the hierarchical structure facilitates easy comparisons between different race/ethnicity measures, regardless of the original wording, format, or source of the item.

Table 6a: Level 1 (OMB+) Taxonomy

Race/EthnIcity	Code	Shorthand
HISPANIC	700	hisp
WHITE	100	white
BLACK	200	black
AMERICAN INDIAN OR ALASKA NATIVE	300	aian
ASIAN	400	asian
NATIVE HAWAIIAN OR OTHER PACIFIC ISLANDER	500	nhopi
SOME OTHER RACE	600	sor

Table 6a outlines the broadest summary level (Level 1) of our race/ethnic taxonomy. The categorization used at this level was chosen to be loosely compatible with the minimum categories spelled out by the OMB standard, with two major differences--we make no distinction between Hispanicity and race, and we include the residual SOR category as a type of emergent social identity. As shown in the table, summary level 1 distinguishes 7 major race/ethnicity groups: the 5 OMB races + Hispanics + SORs. For convenience, we refer to the L1 classification as OMB+.

By reducing all race/ethnic responses to a small, fixed number of categories, even vastly different measures (e.g., open-ended ancestry vs. 5-category school race) can be readily compared at summary level 1.
The trade-off is wasted information, since distinct groups are combined under broad OMB+ categories. To preserve the richness of the data, our hierarchical taxonomy also includes two detailed summary levels. The most refined is summary Level 3, which designates 56 unique national, ethnic, and tribal origins.

As a general rule, we assign an L3 code to any specific group that is either a) listed on the senior survey OR b) written in 15 or more times. Thus, most L3 codes correspond to single, specific origins such as Irish, Puerto Rican, or Cherokee. Note that L3 codes are reserved for those who identify *specific* origins. Respondents who identify only in pan-ethnic or racial terms do not receive an L3 code. Respondents who identity a specific group that is not sufficiently represented in our sample (n<15) may still receive an L3 code, but their responses are pooled. Thus, code 459 ("Other Asians") is the L3 code for those who are Burmese, Pakistani, or any other specific Asian nationality that is too small to receive a unique code. Since the number of detailed groups in our taxonomy is size-delimited, users wishing to apply this coding scheme to other data sources may wind up with more or fewer categories, depending on the number and size of sampled race/ethnic populations.

Intermediate to the 7 OMB+ categories (Level 1) and the 56 detailed groups (Level 3) are 25 summary level 2 groupings. Unlike the L3 codes, which are delimited only by size, there is no single justification for the categorization chosen at summary level 2, and alternate coding schemes are certainly plausible. Our goal was to identify potentially "important" (and empirically testable) cleavages in the OMB+ categories that were appropriate for the UW-BHS population and of analytic interest to investigators. Table 6b details the 25 categories at summary level 2.

Table 6b: Level 1 and Level 2 Taxonomy

Race/EthnIcity	Code	Shorthand
HISPANIC	700	hisp
Mex/Pt.Rican/Cuban	710	mpc
South/Central American	720	s/cam
Other Hispanic	740	oshisp
Non-specific Hispanic	790	nshisp
WHITE	100	white
Western/Northern European	110	weuro
Southern/Eastern European	140	seuro
Non-European Whites	160	owhite
Middle East/N.Africa/SW Asia*	170	menaswa
Brazilian or Portuguese	180	brazport
Non-specific White	190	nswhite
BLACK	200	black
Africa*	210	africa
Carribean/West Indian	240	cawi
Black or African American	290	afam
AMERICAN INDIAN OR ALASKA NATIVE	300	aian
American Indian Tribe	310	tribe
Alaska Native*	380	aknat
Non-specific American Indian	390	nsaian
ASIAN	400	asian
Specific Asian	410	sasian
Non-specific Asian	490	nsasian
NATIVE HAWAIIAN OR OTHER PACIFIC ISLANDER	500	nhopi
Native Hawaiian	510	nhawaiian
Pacific Islander	520	opi
Non-specific PI	590	nspi
SOME OTHER RACE	600	sor
Uncodable	610	uc
Don't Know	620	unknown
Unknown Mixture	630	umx
Refusal	640	ref

As shown, most L2 codes distinguish regional subclasses of the OMB+ categories (Western vs. Eastern European; American Indian vs. Alaska Native; African American vs. West Indian). A few L2 groups are given unique codes solely because they seem out of place within their respective L1 parent categories. Isolating questionable subpopulations in this manner allows for a straightforward assessment of the logic and validity of the OMB taxonomy itself. For example, some L2 groups counted under “white” might differ in ways that warrant a separate OMB+ category (e.g. Middle Easterners), while others might be better classified under a different parent category (Brazilian or Portuguese = Hispanic?).

Summary L2 codes also distinguish respondents by the level of detail they provide on race/ethnicity queries. For example, code 520 (“Pacific Islander”) aggregates

respondents who supply *specific* responses like “Guamanian” and “Samoan” (L3 codes 521 and 522, respectively), while code 590 (“Non-specific PI”) is limited to those who identify only as “Pacific Islander” without giving any additional detail.[7] Notably, since these respondents do not identify a specific origin group, no L3 code is assigned to their identities.

We distinguish between vague and specific responses because we suspect each might be used by different types of respondents. Vague respondents are perhaps more likely to be descendents of older immigrant stock or the offspring of an interracial union. Pooling these respondents with more recent immigrants or persons from non-mixed households risks combining subpopulations that differ on more than just national origin (e.g. time in U.S., language usage, etc.), an assumption that can only be tested if vague and specific responses are disaggregated.

Lastly, we assign codes to SOR responses in an attempt to highlight variation among those who fail to identify any OMB race or ethnicity. That mere fact that respondents refuse to do so is interesting in itself, since every non-indigenous person living in the U.S. has ancestors that migrated from some other continent. While the inclusion of Hispanics as a coequal category should dramatically reduce the number of SOR write-ins, there may still be important differences among this small but racially ambiguous population. As a result, we attempt to distinguish between those who are unable to provide a race/ethnic identity (e.g., unknown, adopted) and those who *refuse* to do so (Human, Martian, etc.).

Our taxonomy uses a three digit hierarchical coding scheme for all populations and subpopulations. Each marked or written identity can have up to three codes (one for each summary level), and the summary level of each code is denoted by the number of trailing zeros. As shown in Table 6b, OMB+ categories (Summary Level 1) are denoted by a one-digit integer ranging from 1-7. Since L1 codes denote broad OMB+ categories, there are two trailing zeros in place of the L2 and L3 codes. Summary Level 2 codes are denoted by two non-zero integers—a shared L1 code and a unique L2 code—followed by a terminal zero in place of the L3 code. Level three codes, in turn, contain all non-zero integers.

Table 6c: Summary Levels 1, 2, and 3 for Hispanics

Race/Ethnicity	Code	Shorthand
HISPANIC	700	hisp
Mex/Pt.Rican/Cuban	710	mpc
Mexican	711	mex
Puerto Rican	715	pr
Cuban	718	cuba
South/Central American	720	s/cam
Panamanian	721	pana
Other S/C American Nation	739	os/cam
Other Hispanic	740	oshisp
Non-specific Hispanic	790	nshisp

Table 6c illustrates this hierarchical structure by listing all three summary levels for Hispanics (L1 code 700). As shown in the table, we’ve coded 4 major Hispanic subgroups at L2: Those who mark or write one of the three listed

[7] For example, if the respondents marked “Other Pacific Islander” but supplied no write-in OR simply wrote “Pacific Islander.”

categories (Mexican, PR, or Cuban), those who write in a South or Central American nation, those who write in other specific Hispanic origins (e.g., Spaniard), and those who identify strictly in pan-ethnic terms like Hispanic or Latino. Unique L3 codes are also included for each of the three listed categories, as well as popular write-ins (e.g. Panamanian).

In addition to substantive race/ethnicity codes, there are several technical codes that distinguish solo from "in combination" identities, as well as unique codes for various race/ethnic combinations. The full taxonomy of codes used in the custom UW-BHS race/ethnicity measures is shown in the Appendix.

Coding Procedures: Overview

Having specified aninclusive race/ethnic taxonomy, the next set of tasks is to encode various race/ethnicity items in the UW-BHS. This process involves a) cleaning the raw responses to all open-ended items, b) encoding all written responses, c) standardizing all written and marked identities, and finally d) integrating marked and written identities into summary measures of race/ethnicity.

The variability of the write-ins necessitated an enormous effort to encode each entry. As shown above, UW-BHS respondents were given multiple open-ended inquiries and allowed to supply an unlimited number of responses to each. Unsurprisingly, respondents articulated their identities in a dizzying number of ways, writing *thousands* of unique identities and permutations, with multiple spellings and syntaxes. So varied were their responses that for many items, the number of unique entries rivaled the number on non-missing responses (see Table 3d).

Coding Procedures: Pre-cleaning Raw Write-in Responses

To save time during the coding process, our first step was a preliminary cleaning of the raw write-ins to minimize the number of distinct entries. This process was automated using a basic find/replace script that looped through each of the raw, open ended write-ins and a) eliminated spelling errors and typos and b) attempted to standardize the syntax of multiple responses. The goal was to reduce the number of unique entries without distorting the respondent's identity in any major substantive way. Table 7a displays a sample of write-in responses before and after cleaning.

Table 7a: Sample open-ended write-ins before and after cleaning

Original Write-in	Clean Version
.25 SWEDISH, .25 SCOTTISH, GERMAN, RUSSIAN	Swedish, Scottish, German, Russian
1/2 AFRICAN AMERICAN AND 1/2 JAPANESE	African American, Japanese
AFREICAN AMERICAN, WHITE	African American, White
AFRICAN AM.	African American
AFRICAN AMERICAN/MEXICAN	African American, Mexican
CAUCASIN	Caucasian
CAUCASION	Caucasian
CEOLE / WEST INDIAN	Creole, West Indian
CHINESE-ESPANISH	Chinese, Spanish
DUTCH & POLISH	Dutch, Polish
DUTCH / GERMAN / FRENCH	Dutch, German, French

As shown, cleaning the write-ins standardizes the way many single and multiple race/ethnic identities are recorded, with only a minor loss of detail (written proportions are not retained, e.g.). In the UW-BHS data file, cleaned versions of the open-ended write in variables (s160 s161 s162 s166 s169) are denoted by the suffix “cln” (s160cln, s161cln, etc.)

Table 7b: Reduction of Unique Write-ins via cleaning

	Number of Unique Entries		
Open-ended Measure	**Original**	**Cleaned**	**% Change**
Ancestry	4052	3580	-11.6%
Primary Identity	1296	1067	-17.7%
Reflected Identity	1176	992	-15.6%
Mother's Ancestry	2412	2051	-15.0%
Father's Ancestry	2239	1885	-15.8%

While the raw entries contain dozens of syntax variations, the cleaned responses are more uniform. This time-saving process reduces the number of unique entries that need to be hand coded, as shown in Table 7b. Indeed, by simply deleting extraneous characters (numbers, percentage signs, trailing/leading spaces etc.), standardizing delimiters (replacing slashes and semi-colons with commas), and running a spell check, we were able to cut down on the number of unique entries by 12-18%. Of course a more aggressive cleaning could have reduced the number of entries even further, but in trial runs, we found that it was easy to distort respondents’ identities by replacing entire strings. For example, many respondents shortened “American” to “Am.” or “Am”, while others included non-ethnic descriptive terms like “not” in their responses. Replacing or deleting all instances of such terms often led to undesirable results: “I am not sure” became “I American” while “Arab, not white” became “Arab, white”. To ensure that the data-cleaning did not alter respondents’ identities, we opted for a more conservative approach and limited our automated editing to spelling and syntax errors.

Coding Procedures: Write-ins

The second step in the coding procedures is the manual coding of written responses. UW-BHS staff painstakingly read, and assigned codes from our race/ethnic taxonomy, to every identity supplied on each of the 10 open-ended race/ethnicity queries on the UW-BHS senior survey: the five write-in spaces on the race/Hispanic Origin questions, plus the five fully open-ended ancestry questions.

To save time, coders used cleaned versions of write-in variables when available, and assigned “shorthand strings” (hisp, brit, viet, etc.) to each identity, rather than keying in the numeric codes directly (See Tables 6x and Appendix for list of shorthand codes).

The race/Hispanic origin questions (which contain listed categories as well as write-in spaces) have a relatively small number of unique write-ins (less than 1000 combined) and were thus entirely hand coded. For the more detailed open-ended questions, we saved time by coding only the entries that were unique to each measure, knowing that many entries would be identical across measures (there are only so many different ways to write an identify, after all). Thus, after coding the race, Hispanic origin, and ancestry responses, we combined all the coded identities into a master codefile. We then queried the primary identity responses,

automatically coding any matches found in the database. Unmatched entries were then coded by hand and added to the database. Identities written on the next open ended questions (e.g. mom' ancestry) were then referenced against this updated database, etc. etc. This iterative process cut the number of novel entries needing manual coding by more than half, and each round resulted in fewer and fewer previously unseen variations.

For each write-in response, we code up to four race/ethnic origins, and each origin is assigned two or three codes--one for each summary level. Thus, every raw write-in response has a minimum of two and a maximum of 12 valid codes. For example, a respondent who writes in "WHITE" would be assigned two shorthand codes, "white" (L1 code 100) and "nswhite" (L2 code 190, non-specific white). A respondent who writes "British" would be coded "white" (L1 code 100), "weuro" (L2 code 110, Western European), and "brit" (L3 code 112, British/English). See Table 6b and the Appendix for full listing of codes.

For respondents who identified multiple origins, the code sequences can be quite long, since each identity is coded up to three times. For example, if a subject reports his father's ancestry as "SAMOAN, CHINESE, GERMAN, AND PUERTO RICAN," he would be assigned the following code sequence:

Level 1	**Level 2**	**Level 3**
nhopi,asian,white,hisp;	opi,sasian,weuro,mpc;	samoa,china,germ,pr

This coding structure provides tremendous flexibility. The hierarchical structure allows us to distinguish respondents at multiple summary levels, while coding the entire string allows us to preserve both the number and sequence of respondents' identities without having to choose one over the other. Under our scheme, those who write in "Asian" or "Vietnamese" can be distinguished from one another, as can respondents who identify as "Black and White" or "White and Black," respectively.

Naturally, the number of unique codes assigned to each response will vary by summary level. Consider the write-in "Afro-American, Jamaican, English, and Romanian," which would be coded:

Level 1	**Level 2**	**Level 3**
black,black,white,white;	afam,cawi,weuro,seuro;	.,jam,brit,roma

This response has several distinct codes at L2 and L3, but only two unique L1 (OMB+) codes.[8] Indeed, it is common for respondents to receive L1 sequences such as "asian, asian" or "white, white, white" if they identify multiple ethnic groups that do not cross OMB+ categories (e.g., "Chinese & Japanese" or "English, German, and Irish"). These groups, while pooled at the OMB+ level, are nonetheless distinguished at more refined summary levels (L2 and L3).

The vast majority of respondent's identities are straightforward and simple to code. Most subjects supplied OMB or OMB-reducible write-ins, and even non-OMB write-ins fell under only a small number of categories (American, Don't know, etc.). Still, a number of written responses were difficult to code, and some required tough judgment calls on the part of the UW-BHS staff.

Generally speaking, we took a cautious approach to ambiguous identities, assigning SOR codes to write-ins like "Creole, biracial, and Jewish." Reasoning that that we

[8] Note that "." signifies missing L3 values (e.g., entries like "Afro-American" and "European American" coded to summary level 2 only).

would have opportunities to obtain valid race/ethnic identities from other survey questions or data sources, we opted to withhold assignment of OMB+ categories unless the identities were fairly unambiguous. The result of this conservative approach is a somewhat inflated count of "SOR" responses on many items.

Table 7c: Examples of problem write-ins and codes

	Shorthand Code		
Write-in	**Level 1**	**Level 2**	**Level 3**
JEWISH	sor	uc	
AMERICAN	sor	ref	amer
HUMAN	sor	ref	amer
ELF	sor	ref	nonsense
STOP ATTACHING RACE TO HOW A PERSON THINKS SICKOS	sor	ref	aver
MIXED	sor	umx	
I WISH I KNEW	sor	unknown	
HEINZ 57	sor	umx	

Of course some write-ins were sufficiently ambiguous that it was impossible to assign an OMB+ identity. Table 7c lists examples of these problem write-ins from s159u (the "some other race" space) as well as the shorthand codes we assigned to them. While none of these identities provide a clear race/ethnic background, there are definitely differences between them, which we attempt to capture with unique "sub-SOR" (L2 and L3) codes. For instance, some responses indicate multiracial/multi-ethnic origins (albeit unspecified ones), while others indicate emerging "Americanized" or "post-racial" identities. Fortunately, these responses are just one of many opportunities for respondents to identify themselves on the senior survey, and since each measure is coded separately, SOR codes on a given measure may be followed (or preceded) by a valid race/ethnic identity on other measures. A handful of respondents, e.g., are coded SOR for write-ins such as "Not sure" or "no primary" on the spaces for AIAN tribes or primary identities, respectively, even though they supplied OMB+ codable information elsewhere.

While our general policy is to avoid "assigning" identities, in a handful of cases we favored a particular interpretation even though an argument could be made for assigning a different category or leaving the identity ambiguous. In most instances, these coding decisions are consistent with those used by the Census Bureau (XXXXX). For example, we took the response "Indian" to mean "American Indian" instead of Asian (though Indian American, Asian Indian, and East Indian are all coded as Asian). Likewise, we interpreted "Hawaiian" to mean Native Hawaiian, and "Indonesian" is treated as an Asian origin, rather than a Pacific Islander one. The responses "Spanish" and "Spaniard" are both coded 700 (Hispanic) at level one, though we count only the former as an indication that the respondent (or his/her recent forbears) emigrated from Spain. Thus entries like "Spaniard" and "from Spain" receive L2 code 740 (Other Specific Hispanic) while "Spanish" (a synonym for Hispanic/Latino, and often listed alongside both in Census queries (see Figure 1)) is assigned code 790 (Non-specific Hispanic).

A number of smaller coding issues warrant brief mention. In distinguishing identities, we tried to ignore redundancies (e.g., "English, British" or "Black, African American" each counted as single identity). Likewise, we attempted to distinguish single identities denoted by compound words (e.g. French Canadian or African American) from identities that suggested multiple ancestries (e.g. French &

Canadian or Nigerian & American), though in some cases, simple comma placement could result in divergent codes (both “Creole” and “French Creole” are coded “SOR” at summary level 1, while “French, Creole” is coded “White, SOR”). We also inferred some identities from non-ethnic terms (slavery references were coded as African American, e.g), and in cases where both specific and ambiguous information were supplied (e.g. “I don’t know, French?”), we coded both, though we tried to place the specific identity *first*. This is one of the only instances in which we ignore the exact ordering of responses, since doing so would force us to code otherwise identical responses differently. For example, rather than coding “Mexican, mostly” and “Mostly Mexican” as two different sequences, we assign the same sequence--“Mexican” and “unknown mixture”--to both. The key advantage of this approach is that is maximizes the amount of ethnic information for analysts who prefer more parsimonious measures of race/ethnicity--in this case, the first identity reported.

Some coding decisions are less defensible than others, and while we attempted to make theoretically informed decisions, we often favored uniformity over a strong theoretical justification. Nowhere was this more apparent than in our decision to apply our race/ethnic taxonomy *uniformly across all measures*. In other words, each write-in is consistently assigned the same code regardless of where it appears. “Black” and “Cherokee” are coded “black, afam” and “aian, tribe, cher,” whether they are written in the Hispanic origin question, the race question, or any of the open-ended questions.

Though elegant in its simplicity, this strategy runs the risk of coding identities in ways that are inconsistent with respondent’s intentions. For example, “Filipino” is coded as an Asian origin even if it is written in response to the Hispanic origin question (under “Other Spanish”). Given the history of Spanish colonization in the Philippines, this decision is potentially problematic, since the respondent might well be of Spanish descent. Still, the alternative--coding Filipino as a Hispanic identity--presents even greater problems. If Filipino counts as a valid Hispanic identity, why not count other seemingly non-Hispanic write-ins such as Korean or Japanese? And if we code Asian identities as valid Hispanic responses, should we code them as something else if written on the “Other Asian” section of the race question?

Rather than engaging in an endless cycle of second-guessing, we chose a uniform taxonomy as the lesser of two evils, since the alternative would have meant creating multiple, item-specific taxonomies. In addition, we provide custom variables that clearly distinguish between the content of written identities and the location (on the questionnaire) in which they are supplied (see below). Researchers wishing to treat every written identity “as is” can simply use the latter measures, if desired.

After the responses to the open-ended write-in spaces were coded, a series of new custom variables was created. For each of the 10 original write-in spaces on the senior survey, there are 16 new variables: One variable with the full, 12-code sequence (all summary levels and origins), three variables with the 4-code sequences for each summary level, and 12 variables with the specific code for each origin and level. Every write-in space has a 3-4 character “root” that is used in naming its custom variables (hsp = Hispanic; ami = American Indian/Alaska Native; asn = Asian, opi = Pacific Islander; sor = Some Other Race; anc = ancestry; prim = primary identity; ref = reflected identity; manc = mother’s ancestry; danc = father’s ancestry). Table 7d displays names and descriptions for all 160 custom variables.

Table 7d: Names and Descriptions of Coded Write-in Variables

	Source of Original Write-in									
Custom Variable Content	Other Hispanic	AIAN Tribe	Other Asian	Other Pacific Islander	Some Other Race	Ancestry	Primary Identity	Reflected Identity	Mother's Ancestry	Father's Ancestry
Full 12-Code sequence	wihsp	wiami	wiasn	wiopi	wisor	wianc	wiprim	wiref	wimanc	widanc
Level 1 sequence	wihspl1	wiamil1	wiasnl1	wiopil1	wisorl1	wiancl1	wipriml1	wirefl1	wimancl1	widancl1
Level 2 sequence	wihspl2	wiamil2	wiasnl2	wiopil2	wisorl2	wiancl2	wipriml2	wirefl2	wimancl2	widancl2
Level 3 sequence	wihspl3	wiamil3	wiasnl3	wiopil3	wisorl3	wiancl3	wipriml3	wirefl3	wimancl3	widancl3
1st L1 Code	wihspcd1	wiamicd1	wiasncd1	wiopicd1	wisorcd1	wianccd1	wiprimcd1	wirefcd1	wimanccd1	widanccd1
2nd L1 Code	wihspcd2	wiamicd2	wiasncd2	wiopicd2	wisorcd2	wianccd2	wiprimcd2	wirefcd2	wimanccd2	widanccd2
3rd L1 Code	wihspcd3	wiamicd3	wiasncd3	wiopicd3	wisorcd3	wianccd3	wiprimcd3	wirefcd3	wimanccd3	widanccd3
4th L1 Code	wihspcd4	wiamicd4	wiasncd4	wiopicd4	wisorcd4	wianccd4	wiprimcd4	wirefcd4	wimanccd4	widanccd4
1st L2 Code	wihspcd5	wiamicd5	wiasncd5	wiopicd5	wisorcd5	wianccd5	wiprimcd5	wirefcd5	wimanccd5	widanccd5
2nd L2 Code	wihspcd6	wiamicd6	wiasncd6	wiopicd6	wisorcd6	wianccd6	wiprimcd6	wirefcd6	wimanccd6	widanccd6
3rd L2 Code	wihspcd7	wiamicd7	wiasncd7	wiopicd7	wisorcd7	wianccd7	wiprimcd7	wirefcd7	wimanccd7	widanccd7
4th L2 Code	wihspcd8	wiamicd8	wiasncd8	wiopicd8	wisorcd8	wianccd8	wiprimcd8	wirefcd8	wimanccd8	widanccd8
1st L3 Code	wihspcd9	wiamicd9	wiasncd9	wiopicd9	wisorcd9	wianccd9	wiprimcd9	wirefcd9	wimanccd9	widanccd9
2nd L3 Code	wihspcd10	wiamicd10	wiasncd10	wiopicd10	wisorcd10	wianccd10	wiprimcd10	wirefcd10	wimanccd10	widanccd10
3rd L3 Code	wihspcd11	wiamicd11	wiasncd11	wiopicd11	wisorcd11	wianccd11	wiprimcd11	wirefcd11	wimanccd11	widanccd11
4th L3 Code	wihspcd12	wiamicd12	wiasncd12	wiopicd12	wisorcd12	wianccd12	wiprimcd12	wirefcd12	wimanccd12	widanccd12

The value in coding identities at multiple summary levels is the flexibility it affords data users. While the raw (or only lightly cleaned) write-ins are far too detailed to be used in statistical analysis, collapsing these rich measures down to six or seven OMB-category variables defeats the purpose of collecting open-ended identities in the first place. Our approach provides the best of both worlds: retaining nearly all the richness of the original responses, including the number and order of identities supplied, while providing convenient summary categories for groups that are too small to be used for statistical inference.

Table 7e. Number of Unique Write-in Categories for Original and Custom Variables

	Source of Original Write-in									
Variable Content	Other Hispanic	AIAN Tribe	Other Asian	Other Pacific Islander	Some Other Race	Ancestry	Primary Identity	Reflected Identity	Mother's Ancestry	Father's Ancestry
Original Write-in (uncoded)	170	175	61	33	291	4052	1296	1176	2412	2239
All origins, all levels (full 12-code sequence)	80	45	40	18	161	2872	542	441	981	1277
All origins, Level 3 sequence only	40	26	32	12	110	2261	320	240	749	941
All origins, Level 2 sequence only	65	32	22	15	107	921	261	270	345	448
All origins, Level 1 sequence only	34	19	12	11	42	315	99	127	128	158
1st origin, L3 Code	19	14	23	8	40	56	55	51	54	56
1st origin, L2 Code	22	19	15	10	22	25	25	25	25	25
1st origin, L1 Code	7	7	6	6	7	7	7	7	7	7

Table 7e illustrates the flexibility of this coding scheme. Rows differentiate variables by summary level and number of origins coded, while columns represent the original source of the write-in on the senior survey (other Hispanics, Ancestry, etc.). Each cell is a frequency count of unique categories for each custom variable.

As shown in the table, coding the write-ins greatly reduces the number of race/ethnic categories while preserving much of the complexity and nuance of the original responses. For those whose identities are detailed enough to warrant level 3 codes (specific nations, tribes, and ethnicities), dozens of unique identities and combinations remain available for analysis, though even our most detailed coding of the (up to) four-origin sequences reduces the number of unique categories by 40-80%, relative to the original, unedited variables. The L2- and L1-coded sequences, which are available for every UW-BHS respondent, further simplify the data while retaining both the number and order of identities written.[9]

Researchers who wish to avoid the complexity of multiple origins have several options. Perhaps the best solution is to eschew “processing” altogether and use the respondent’s primary identity (s161) to resolve such cases. If focusing on a specific questionnaire item, however, or in instances where primary identity may be missing or ambiguous, users may focus on the *first origin reported*. Even this simplified approach provides three levels of detail, with the number of unique categories bound by the number of groups coded at each summary level. At the highest level of detail (L3), there are up to 55 unique race/ethnic identities. Users preferring a highly condensed classification can utilize the “first origin, L1-coded” variables, which reduce every original write-in to just seven OMB+ categories.

[9] Some variables have more unique L2 sequences than L3 sequences since some respondents identities’ are not detailed enough to be coded at summary level 3.

The “first origins” approach provides an elegant reduction of open-ended race/ethnicity queries, though data cannot be reduced without making assumptions. Assigning only the first identity reported risks pooling potentially distinct respondents or distinguishing potentially similar ones. It assumes, for instance, that the first OMB identity reported is the most salient. Thus “black and white” and “white and black” would have different first origins. It further assumes that OMB identities are more salient than non-OMB identities, even if the latter are reported first (recall that OMB identities are coded ahead of non-OMB identities in instances where both occur). Thus, the write-in “Don’t know, maybe Irish” would code Irish as the first origin, even if this ordering seems highly suspect. Users should therefore exercise caution when relying on respondent’s first origins. In most cases, primary race should provide a less biased solution.

Standardizing Written and Marked Identities: Location vs. Content

After encoding the write-ins, the next step is to combine written identities with those that are marked or circled. But since the race/ethnicity measures on the senior survey vary in format (open-ended vs. closed, “choose one” vs. “choose all that apply), responses must be standardized before they can be combined. The most straightforward approach is to create dichotomous indicators for each “identity”, marked or written. Each indicator would then be coded 0 or 1, depending on whether the respondent marked/wrote that identity or not.

Though simple in principle, the complexity of the race/ethnicity data in the UW-BHS presents challenges even for this seemingly procedural step. Multiple measures beg the question of whether a single set of dichotomous indicators is sufficient, or if separate indicators should be coded for each measure. Neither approach is without problems. On one hand, coding separate indicators for each survey question, or worse yet, *each write-in space*, would greatly expand the number of variables needed (88 coded identities by 10 write-in spaces = 880 indicator variables) and create a dizzying number of identity-item combinations.

On the other hand, coding a single, grand indicator (e.g., “Any Native American Ancestry”) risks conflating strong identities (AIAN alone by race) with more vestigial ones (“tiny bit of Indian” as mother’s fourth ancestry). Respondents could even be coded as having identities that are only attributed to them by others (self-reported black whose reflected race is white), or that they claim by mistake (writing “Irish” in the space for AIAN tribes). For these reasons, one of the more difficult issues in standardizing identities lies in how, or even *whether*, to resolve those that are potentially misplaced. As shown in Section II, many respondents seem to misinterpret race/ethnicity questions and categories (particularly residual categories). Whether they write Hispanic origins in the space designated for “other race,” non-Hispanic origins in the space designated for “other Hispanics,” or Pacific Islander groups in the space intended for “other Asians,” a non-trivial number of persons supplied contradictory information by writing one identity in places *intended* for others.

These examples all highlight the need to distinguish between the *location* and *content* of race/ethnic identities in social research: the raw mechanics of identities vs. the presumed meaning. Location refers to the section of the questionnaire in which various race/ethnic identities are queried. Locations can be whole questions (Hispanic origin in Q. 158, race in Q. 159, etc.) or any parts thereof (“black” checkbox or “Other Spanish…” write-in space). Content, by contrast, denotes the substance of the identity provided in each location. Ideally, location and content would have a 1:1 correspondence, with members of groups X and Y responding only in the sections designated for groups X and Y, respectively. In practice, there are instances in which respondents identify *as* Group X *in* section Y

and vice versa. This mismatch presents a problem, since there are now two identities from which to choose. Consider a respondent who writes "Irish" as their Hispanic origin. If we privilege the mere act of identifying *in the Hispanic Origin section*, we would code this person as Hispanic, along with anyone else who circles or supplies a write-in under "Other Spanish…" If, on the other hand, we code the *substance of the identity* itself, "Irish" and similar non-Hispanic write-ins would not be counted as Hispanic, even if they happen to be written in a location intended for Hispanics.

Rather than choosing one interpretation over the other, we construct dichotomous indicators for both types of identification. Thus, each race/ethnic group has a set of A) content indicators that distinguish whether the respondent marked or wrote *that particular group*, and B) location indicators that distinguish whether the respondent marked or wrote *anything in the space(s) designated for that group*.

Both the location and content indicator variables combine a four letter prefix with a race/ethnic "root word" (wht, blk, ami, asn, opi, sor, hsp, etc.). Location indicators are prefixed with the characters "**idin**" (identifies in…), while content indicators are prefixed with the characters "**idas**" (identifies as…). While the root words are the same in both sets of variables, the interpretations are different. A person who identifies "as Asian" (**idasasn** = 1) either writes or marks an Asian identity (e.g., Japanese). A person who identifies "in Asian" (**idinasn** = 1) simply provides a response (Asian or otherwise) in the "Asian section" of the survey.

Location indictors are available only for OMB+ groups and are assigned positive values if the respondent either A) marks one or more categories associated with that group or B) supplies a write-in *of any kind* in the optional space for that group. Table 8a lists source variables for each constructed location indicator. As shown in the table, the "location" of each OMB+ group corresponds to the *section*

Table 8a. Name and Description of OMB+ location indicator variables

Location Indicator Variable	OMB+ Group	Original Source Variables (1) (2)			
idinwht	White	s159a			
idinblk	Black	s159b			
idinhsp (3)	Hispanic	s158a-b			
idinami	AIAN	s159c-d			
idinasn	Asian	s159e-n			
idinopi	NHOPI	s159o-s			
idinsor	Some Other Race	s159u-t			
(1) For dichtomous source variables (i.e., markable race categories), Location indicators take a value of 1 if any source variable equals 1					
(2) For open-ended (write-in) source variables, Location indicators take a value of 1 if the source variable is non-missing (i.e. any write-in supplied)					
(3) Coded 1 for responses other than a standalone "non-Hispanic" on s158a					

of the race/Hispanicity question in which each group is queried, as measured by the original source variables (see Sections I and II). Some groups, like whites and blacks, are enumerated by a single listed category and have just one dichotomous source variable. Thus, the white and black "sections" are simply the first two responses on the race question.

Other OMB groups are designated combinations of listed categories and write-in spaces. AIANs have one of each—the listed category "American Indian or Alaska Native), and a write-in space for detailed tribes. The Asian section spans several listed origin groups as well as a residual category ("Other Asian") with a write-in space. Any respondent who either circles an Asian category OR supplies a write-in under "Other Asian" is coded 1 on **idinasn**. Hispanic identities are measured in the Hispanic origin question (s158), and while both Hispanics and non-Hispanics are instructed to fill out the question, we do not consider the standalone response "non-Hispanic" as part of the Hispanic section. Thus someone, who marks "non-Hispanic" (code 1 in Table 1a) and provides no write-in would be coded 0 on the location indicator **idinhsp**.

Table 8b: Race/ethnicity Sections Used

			Percent	
OMB+ Group	**Location Indicator Variable**	**Freq (1)**	**Of Total (2)**	**Of UW-BHS Sample**
White	idinwht	6091	52%	63%
Black	idinblk	1359	12%	14%
AIAN	idinami	511	4%	5%
Asian	idinasn	1759	15%	18%
NHOPI	idinopi	290	2%	3%
Some Other Race	idinsor	712	6%	7%
Hispanic	idinhsp	937	8%	10%
	Total (2)	**11659**	**100%**	**121%**
	UW-BHS Sample Size	**9659**		

(1) Number of respondents who used each section

(2) Exceeds sample size since respondents could use multiple sections

Table 8b tabulates the number of respondents who used each race/ethnic location in the survey. Note that counts of respondents who used the white and black sections are identical to the counts from the original source variables, **s159a** and **s159b** (See section I).[10] This is expected, since marking the categories is the only way to identify with these groups,[11] and each group has just one category. Counts of respondents who used other sections (AIAN, Asian, etc.) are logically higher, since respondents could either mark a category or write something in. 712 respondents, for example, either circled SOR or wrote something underneath it. Notably, since respondents can identify in as many sections as they wish, the sum of positively scored content indicators exceeds the number of UW-BHS respondents. On average, each respondent identified in 1.2 race/ethnic sections. This should not be interpreted as an estimate of ethnic/racial mixture, however, since the content of the write-ins is not considered.

[10] While counts of positive scores are identical, counts of zeroes differ from the original source variables, since the content indicators also code missing values (i.e. marking or writing no categories anywhere in the race question).

[11] While white and black respondents are free to identify as such in other sections and questions, these two lines on the race question are the only spaces on the survey *specifically* reserved for these two groups

The senior survey lists several detailed origin groups as categories. While it is possible to code location indicators for sub-OMB+ groups, only those that are listed by name have exclusive "locations" on the survey (Japanese, Mexican, Samoan, etc.). Groups that are not listed (Panamanian, Cherokee, Human, etc.) can only be counted by disaggregating the write-in responses. Because the location indictors consider only the presence/absence, and not the substance, of responses in particular sections, they are unsuitable for more detailed levels of analysis.

To account for this limitation, we also create a series of content indicators that measure substantive claims of various race/ethnic identities. Naturally, since respondents can identify in multiple questions and at multiple levels of detail, the number of content indicators exceeds the number of location indicators. First, unlike marked identities, which are constrained by the wording of the listed categories, written identities can be at any level of detail. Thus, for each race/ethnic group in our taxonomy, separate content indicators are created for a) the most detailed (L2 or L3) code available and b) the OMB+ (L1) code. Respondents are then scored for each identity and summary level they report. For example, a respondent who marks or writes "Japanese" would be coded 1 on the variables **idasasn** and **idasjapa**, while a respondent who writes "Peruvian" would be coded 1 on the variables **idashsp** and **idasscam** (South/Central American Nation). A respondent who identifies only as "Asian" (either by writing "Asian" or marking "Other Asian…" and supplying no write-in) would be coded 1 on the variables **idasasn** and **idasgasn** (general Asian).

Second, while each OMB+ group has just one designated section in the survey, respondents are free to write their identities in any section(s) they wish. As discussed above, coding separate indicators for each survey item would require an ungainly number of new variables. But the alternative, coding a single content indicator (e.g., "Black on any item"), risks conflating strong identities with secondary, even symbolic ones. To preserve these important distinctions while keeping the number of variables manageable, we create two sets of content indicators: one for identities claimed on any survey question (except reflected race), and another for identities claimed only on the race and Hispanic origin questions. The latter are indexed by the suffix "15x" (questions 158 and 159 only), while the former use the suffix "1xx" (any race/ethnicity question in the 150s and 160s). Thus, someone who identifies as Hispanic on the race/Hispanic origin question would be coded 1 on the variable **idashsp15x.** Someone who identifies as Black on any question would be coded 1 on the variable **idasblk1xx.**

Table 8c provides a partial listing of the content indicator variables for each L2-L3 race/ethnic group in the taxonomy (coded thus far[12]). Note that in addition to creating indicators for specific L2/L3 identities, we also include indicators for "generic" (or general) responses (e.g., ghsp = General Hispanic), coded 1 for respondents who identify only in vague racial or pan-ethnic terms.

Table 8c

Partial Variable Description	Content Indicator varname	
	In Race/Hispanic Origin	In any Survey Question
ID's as Mexican	idasmexi15x	idasmexi1xx
ID's as Puerto Rican	idasprcn15x	idasprcn1xx
ID's as Cuban	idascuba15x	idascuba1xx

[12] Note: As of 11/2009, sub-OMB content indicators are available only for Hispanic, Asian, AIAN, NHOPI, and SOR sub-groups that meet critical size thresholds (white and black subgroups forthcoming).

ID's a S/C American Nation	idasscam15x	idasscam1xx
ID's as Panamanian	idaspana15x	idaspana1xx
ID's another S/C Amer. Nation	idasotsc15x	idasotsc1xx
ID's another Hispanic Origin	idasohsp15x	idasohsp1xx
ID's simply as 'Hispanic'	idasghsp15x	idasghsp1xx
ID's as Brazilian or Portuguese	idasbzpt15x	idasbzpt1xx
ID's as Cherokee	idascher15x	idascher1xx
ID's as Blackfoot	idasbkft15x	idasbkft1xx
ID's some other Tribe	idasotrb15x	idasotrb1xx
ID's as Alaska Native	idasaknv15x	idasaknv1xx
ID's simply as 'Indian' or 'Native Amer.'	idasgami15x	idasgami1xx
ID's as Asian Indian	idasasin15x	idasasin1xx
ID's as Cambodian	idascamb15x	idascamb1xx
ID's as Chinese	idaschin15x	idaschin1xx
ID's as Filipino	idasfili15x	idasfili1xx
ID's as Japanese	idasjapa15x	idasjapa1xx
ID's as Korean	idaskore15x	idaskore1xx
ID's as Laotion	idasloat15x	idasloat1xx
ID's as Vietnamese	idasviet15x	idasviet1xx
ID's as Indonesian	idasindo15x	idasindo1xx
ID's as Thai	idasthai15x	idasthai1xx
ID's some other Asian Origin	idasoasn15x	idasoasn1xx
ID's simply as 'Asian'	idasgasn15x	idasgasn1xx
ID's as Nat. Hawaiian	idasnhaw15x	idasnhaw1xx
ID's as Guamanian	idasgmch15x	idasgmch1xx
ID's as Samoan	idassamo15x	idassamo1xx
ID's Other PI Origin	idasospi15x	idasospi1xx
ID's simply as 'NHOPI'	idasgopi15x	idasgopi1xx
Supplies Uncodable response (usually religion)	idasuncd15x	idasuncd1xx
Writes 'Don't Know' or 'Unknown'	idasdtno15x	idasdtno1xx
Writes 'Mixed' or similar	idasumix15x	idasumix1xx
Writes 'American' or 'Human'	idasamer15x	idasamer1xx
Supplies Non-sensical response (Martian, elf, e.g.)	idasjoke15x	idasjoke1xx
Expresses Aversion toward R/E questions	idasaver15x	idasaver1xx

Which indicators should be used?

The location and content indicators serve distinct yet overlapping goals, and each has unique strengths and limitations. Most respondents will be scored the same way

on each set of indicators. This is particularly true for racial categories that are not followed by a write-in. Here respondents have just two choices: Circle the category or skip it. Thus, a respondent who circles white identifies *in the white section* and also identifies *as white,* and would thus be coded 1 on both dummy indicators (**idinwht** and **idashwt**). The only divergence would be a slightly higher count in the content indicators, due to additional whites being identified via the content of written responses to other sections (e.g. writing "Irish" under "other race"). Likewise, a respondent who marks one or more specific Asian, NHOPI, or Hispanic origins would also be scored identically on both sets of indicators. Of course, the mere presence of a marked category does not ensure that the respondent meant to identify with that group. He/she might well have made an error. But in the absence of competing information (like a contradictory write-in), there is simply no way to qualify or discredit a specific, marked origin.

In other instances, competing information *is* available—namely, when respondents use the write-in spaces. In such instances, it makes sense to consider the content of the written information, as this is usually where the two sets of indicators diverge. The location indicators give no consideration to the substance of written identities. All write-in responses to a given race/ethnicity section are given equal weight; the sole criterion for inclusion is that the space not be left blank. Content indicators, on the other hand, rely on the taxonomic codes assigned to each written identity. For a written response to be coded as a positive indicator of membership within a race/ethnic group, it must logically interpretable as such. Given the divergent approaches each set of indicators takes toward the write-ins, each has the potential to score written identities differently.

For most written identities, the two indicators will not vary. A write-in of "Panamanian" under "Other Spanish…" is both an identification in the Hispanic section and a substantive identification *as* Hispanic. But a write-in of "Italian" in the same space would count only as identification in the Hispanic section, and *not* as a Hispanic identity, per se. This distinction is important, since the content indicators will override a marked identity if the accompanying write-in supplies contradictory information. Our reasoning is that it is easier to mistakenly mark an identity through haste or carelessness than it is to mistakenly write one in. Respondents may inadvertently circle categories they don't fully understand, particularly residual categories ("Other Spanish, Other Race, etc.), which they may (mis)interpret as a suitable location to identify themselves on a given question. Identities that respondents take the time to write out, by contrast, are likely more deliberate, purposeful claims about how they see themselves.

Each coding alternative has implications for population measurement. Location indicators privilege respondents' raw, unedited information, taking every identity "as is," even at the risk of measurement errors. Content indicators, by contrast, ignore identities that seem to have been marked or written in error or jest, even at the risk of invalidating or disqualifying some self-reported responses.

Table 8d compares counts of each race/ethnic population using respondents who identify in each group's section with counts of those who identify as each group in any (15x) section.[13] This comparison provides an estimate of the extent to which the location indicators undercount (or overcount) each race/ethnic population, relative to the content indicators. Note that these are estimates of net undercount, since out-group responses in an in-group's location (e.g. writing non-Asian origins in

[13] To ensure an apples to apples comparison, the 15x content indicators are used

the Asian section) must be weighed against in-group responses in all out-groups' locations (e.g. writing Asian origins in non-Asian sections).

Results show wide variation in the direction and magnitude of the divergence between the two sets of measures. Relative to the substantive indicators, the raw location indicators seem to *underestimate* the number of whites and blacks and *overestimate* all other groups. But these results owe, in part, to differences in how each group is measured. Whites and blacks have no optional write-ins, so there is no way to identify, or *correct*, responses that might have been made in error. For this reason, white and black marked responses are treated "as is."

All other OMB+ groups have write-in spaces, so "false positives" can be identified and removed by analyzing the content of the identities written. Thus, for each race/ethnic location with a write-in, the number of internally consistent responses (content indicator = 1) should be lower than the number of total responses, since the latter sum includes errors. Adjusting for content need not reduce the size of a group, since it is also possible to pick up "false negatives" when respondents claim a particular identity in sections designated for other identities (see examples in Section II). This is why groups without write-ins (whites, blacks) are always larger after adjusting for content. Whites and blacks can *only* pick up cases; they never lose any, since there are no white/black write-in sections in which non-whites and non-blacks might mistakenly identify.

This constraint also ensures that non-black minorities, as a whole, will be overcounted by the location indicators, since these are the locations on the survey in which misplaced white- and black- coded identities can be written. Individual groups have no such constraints, so the variation between the content- and location-defined counts can be interpreted in substantive ways. For Asians and AIANs, Table 8d shows that there is very little aggregate difference. This does *not* imply that AIANs and Asians always identity in their respective sections, or that other groups don't mistakenly use the AIAN/Asian sections. It simply means that the two types of errors are offsetting. For Hispanics and NHOPIs, the gaps are larger and positive. Thus, while Hispanics and NHOPIs doubtless have some members that mistakenly identity in other sections of the survey, the more common error is for out-group persons to mistakenly identify in Hispanic/NHOPI sections. As a result, if we treat the mere act of writing something in the NHOPI or Hispanic sections as valid identity claims, we will overcount the number of substantive NHOPIs and Hispanics by 9% and 12%, respectively.

Table 8d: Counts of Race/ethnic Identification by Location and Content

	Number that identifies		
OMB+ Group	As the Group	In the Section	Bias (1)
White	6300	6091	-3%
Black	1387	1359	-2%
AIAN	500	511	2%
Asian	1757	1759	0%
NHOPI	267	290	9%
Some Other Race	136	712	424%
Hispanic	833	937	12%

(1) Percent by which location indicators overstate group size

Finally, while the choice between the content and location indicators is clear, the choice *of* content indicators is somewhat less so. The 1xx and 15x indicators both incorporate information from multiple sections of the survey, and both consider the substance of written identities as well as marked ones. In addition, neither set of indicators incorporates reflected race responses, since the latter may contain identities that respondents would not claim for themselves.

The 15x indicators are derived from responses to the race and Hispanic origin variables only. As noted above (see Table 5 and surrounding discussion), the empirical and conceptual overlap between race and Hispanic origin led us to combine these two measures into an integrated race/ethnic taxonomy. We do the same for the content indicators, since neither race nor Hispanic origin alone does a suitable job of distinguishing UW-BHS respondents. Hispanic origin (s158) only lists Hispanic groups, while race (s159) only lists Non-Hispanic groups. Yet Hispanics and non-Hispanics are instructed to fill out both questions. Lacking specific in-group categories with which to identify, it is unsurprising that each group either skips what they perceive to be an irrelevant question, or spills over into the residual categories for each respective question—Hispanics in “other race,” non-Hispanic in “other Spanish origin.” The end result is two questions that measure opposite ends of the population. Combining the two is both logical and necessary. Thus, rather than coding separate content indicators for race and Hispanic origin, we treat both as two halves of a single question about race/ethnic identity, and code a single “15x” (questions 158 and 159) indicator for all respondents.

While both questions combined do a better job than either question alone, by counting only race/Hispanic origin, the 15x indicators give no weight to identities that are potentially more vestigial or symbolic (mother’s third ancestry, e.g.) than the “hard” inquiries on race/Hispanicity. As such, the 15x indicators sacrifice both coverage and nuance for precision by omitting respondents who may have overlooked or deliberately skipped these questions.[14] The 1xx indicators cover everyone who supplies information about race/ethnicity/ancestry, though not without making the opposite tradeoff—increasing coverage while reducing precision. For example, only those who mark or write “American Indian" as a race would be coded as AIAN on the 15x indicators. Contrast this to the 1xx indicators, in which the respondent need mention only a trace of AIAN ancestry to receive the same code.

Since the 1xx indicators are a superset of the 15x indicators, positively scored 1xx counts are always equal to or greater than those from the 15x indicators for the same race/ethnic group; likewise, the difference between the 15x and 1xx indicators provides a count of those who identify each group on the open-ended “16x” questions but not in the race/Hispanic origin questions

Coding race/ethnic responses from other data sources

Neither supplemental data source contains a large number of write-ins, and neither lists race/ethnic categories at the sub-OMB level. The administrative data contain no write-ins and are already limited to OMB categories (see Table 4a). Thus, the admin data are simply recoded to summary level 1 categories.

The EMS does contain an open-ended SOR category with space for raters to type in identities other than those listed. As shown in Table 4b, this occurs in just 84 of

nearly 30,000 total viewings (0.3%). After applying the same coding procedures used on the senior survey, just 10 ratings could not be coded to an OMB category.

IV. Summary Measures of Race/Ethnicity

Having coded the write-ins and created a series of dichotomous indicators for both the content and location of race/ethnic identities, we now proceed to the task of combining various identities and combinations into coherent, summary measures, that assign is unique identity to each respondent. Given the multiple sets of dummy indicators, there are several classes of summary race/ethnicity measures that have been developed. The broadest distinction is between those derived from the content and location indicators, respectively. Within the content indicators, which we treat as the "cleaner" of the two measures, there are additional distinctions by A) the number of categories and combinations coded, B) rules for allocating ambiguous cases, and C) level of detail (racial origins, ethnic origins, or both).

We start by looking at the location-based measures, since these are far simpler to work with.

IV.A.: Summary Location Measures

All summary location measures are permutations of the dichotomous location indicator variables ("**idin**…", which in turn are closely derived from the raw, unedited responses to the Hispanic origin (s158) and race (s159) questions (see section . Since we are dealing only with the location in which identities are supplied, the content of written responses is not considered here. "idin*" location indicators shown in Table 8b. The goal is simply to reorganize the raw race/Hispanicity responses so that they sum to 100% of the sample size, rather than summing to the total number of identities claimed, which exceeds the sample size since respondents may claim more than one.

While the location measures ignore content, some minor editing and pooling is required to reorganize the data so that identities and persons have a 1:1 ratio. We start with the Hispanic origin question, which is recoded into the summary location variables **s158ab** and **s158ax**.

Table 9a: Variable s158ab: s158ax with all write-ins coded to "Other Hispanic"

Code	Group	N	Percent
99	Non-Hispanic	8118	84.05
711	Mexican	413	4.28
715	Puerto Rican	125	1.29
718	Cuban	25	0.26
740	Other Hispanic	374	3.87
.	MISSING	603	6.24
		9658	**100**

s158ax is identical to the original **s158a** except that it recodes the two cases who circled multiple categories (see Table 1a) from fractional codes (e.g. 2.5) to "other Hispanic…." **s158ab** further transforms **s158ax** by combining information from **s158a** (marked categories) and **s158b (**written information) to provide a richer account of the nuanced ways in which respondents fill out the H/O question (see Table 1c). In addition to persons who marked a Hispanic category in **s158a**, **s158ab** additionally codes all persons who supplied a write-in under "Other Spanish…" as

members of the "Other Hispanic" category. Notably, this includes persons who wrote something on the Hispanic Origin question but did not mark a Hispanic Category. All told, **s158ab** adds 38 new "write-in only" Hispanics, including 28 who failed to mark anything (and thus are coded missing on s158a and s158ax), and another 10 who marked non-Hispanic only to then write something under "other Spanish…" The four Hispanic subgroups in s158ab sum to 937, the same number coded 1 on the dichotomous location variable **idinhsp.**

Table 9b: Variable idin009159: A 9-group Summary of Raw (uncleaned) Responses to the UW-BHS Race Question

Code	Group	N	Percent
101	White Alone	5063	52.42
201	Black Alone	870	9.01
301	AIAN Alone	96	0.99
401	Asian Alone	1256	13
501	NHOPI Alone	152	1.57
601	SOR Alone	373	3.86
606	SOR In Combo	339	3.51
10000	MULTIRACIAL	959	9.93
.	MISSING	550	5.69
		9658	**100**

The equivalent location-based summary of race is the custom variable **idin009159**, shown in Table 9b.[15] **idin009159** is a permutation of raw marked & written responses to the race question. The source variables are the location indicators shown in Table 8b. Note, since **idin009159** is derived only from question 159 (race), no Hispanic origin information is included. As shown in Table 9b, **idin009159** codes each monoracial respondent to the single OMB+ category that they marked and/or supplied a write-in beneath. Sub-OMB identities are ignored, so a respondent who marked Chinese plus Japanese would be coded Asian alone. Only those who circled categories that fall under two or more major racial groups are counted as multiracial. **idin009159** also distinguishes those who marked (or wrote something under) SOR alone, and those who marked/wrote an SOR and also used another (standard) section of the race question.

Table 9c: Variable idin00815x: An 8-group Summary of Raw Race & Hispanic Origin Responses

Code	Group	N	Percent
01	White Alone	4980	52%
201	Black Alone	832	9%
301	AIAN Alone	72	1%
401	Asian Alone	1215	13%
501	NHOPI Alone	137	1%
601	SOR Alone	175	2%
709	Hispanic of Any Race	937	10%
10000	MULTIRACIAL	951	10%
.	MISSING	359	4%
		9658	**100%**

[15] There is also an 8-category version (**idin008159**) that pools the two types of SOR responses (alone plus in combination).

The final summary location variable is **idin00815x**, shown in Table 9c, which combines information from Hispanic origin (**s158ab)** and race (**idin009159).** **idin00815x** differs from **idin009159,** which uses race only, in two important ways. First, it pools the 937 respondents who supplied a "positive" response to the Hispanic Origin question and codes them as a coequal category to the other racial groups. Thus the standard racial groups in Table 9c (including multis) are limited to self-reported non-Hispanics (**idinhsp**=0). Second, **idin00815x** privileges OMB categories over SOR responses when both are supplied. Thus, if a respondent marked white but also wrote something under SOR, he is coded simply as white in **idin00815x**. Once again, persons who marked (or wrote something under) two or more major OMB racial categories are coded as multiracial.

The summary *location* race/ethnicity measures provide a valuable "first pass" at describing the diversity of the UW-BHS sample. These measures have the same pros and cons as the dichotomous location indicators on which they are based. On one hand, users who desire an unedited, parsimonious coding of the marked and written responses to race and Hispanic origin questions should find measures like **idin00815x** very useful. Location-based measures rely on very few assumptions, and make no effort to interpret (or discredit) the content of written responses. With the exception of privileging Hispanic origin over race, and ignoring SORs that follow standard categories, **idin00815x** summarizes respondents' identities more or less "as is." This simplicity comes with an obvious trade-off, however, as all the ambiguity (and error) of the raw, unedited responses are also passed along "as is."

Written responses are particularly problematic in this regard, as the content of the write-ins is often incompatible with the OMB category under which it is supplied. For example, a person who wrote "German" on the space for Native American tribes is coded identically to a person who wrote "Navajo." Both are AIAN via the *location* on the race question in which they supplied a response, even though their identities are substantively quite different. Likewise, a write in of "I'm black" under Hispanic origin is coded identically to a write-in of "Columbian," since both responses were supplied on the "other Spanish…" line. Obtaining "cleaner" measures of race/ethnic identity requires that we look at *how*, rather than where, respondents identify.

For that we turn to summary race/ethnicity measures derived from the content based indicator variables, which code written responses into substantive race/ethnic categories, wherever plausible. Rather than coding a location (e.g. "Other Spanish…") regardless of what's written, content-based measures code the write-in itself, regardless of its location. Content-based measures encompass additional identities such as ethnic/tribal groups, and some reference multiple race/ethnicity questions, rather than just race and Hispanic origin alone. As a result, the content-based measures are more numerous and more complex than their location-based equivalents. Most of the variables presented in the following sections derived from the Race and Hispanic origin questions, though many draw upon additional UW-BHS items to minimize non-response and/or resolve ambiguous cases.

IV.B: Basic Content-based summary variables: OMB categories only

The first set of summary measures is limited to major OMB+ categories and combinations. The most basic is **r15xgrpall**, an eight-category permutation of the content indicator variables from the "15x" section of the survey. It is reasonable to view **r15xgrpall** as a "cleaned" version of **idin00815x**. Like that variable, **r15xgrpall** is derived solely from race and Hispanic origin (hence the identical counts of missing values), and assigns each respondent into A) a single OMB+ race,

B) Hispanic of any race(s), or C) Non-Hispanic of two of more OMB Races. The variable is named **r15xgrpall** because it contains **r**acial identities from questions 158 and 159 (**15x**) and **gr**ou**p**s **all** Hispanics and multiracial respondents.

While the categories themselves are unchanged, the use of coded indicator variables changes the meaning of group membership. For example, the SOR residual category is now interpretable as an actual ambiguous (non-OMB codable) identity like "Creole" or "Human," rather than a simple indicator that the respondent wrote something in the SOR space (e.g. Irish). Table 9 presents **r15xgrpall** side by side with **idin00815x**, and illustrates the degree of bias in the raw, unedited race responses.

Table 9d: Variable r15xgrpall: Cleaned Non-Hisp Single Race, Hispanic of Any Race, or Pooled Multiracial

		r15xgrpall (cleaned)		idin00815x (raw)		Bias in
Code	**Group**	**N**	**Percent**	**N**	**Percent**	**raw data**
101	White Alone	5152	53%	4980	52%	-3%
201	Black Alone	843	9%	832	9%	-1%
301	AIAN Alone	65	1%	72	1%	11%
401	Asian Alone	1232	13%	1215	13%	-1%
501	NHOPI Alone	124	1%	137	1%	10%
601	SOR Alone	46	0%	175	2%	280%
709	Hisp, Any Races	833	9%	937	10%	12%
18030	2+ Non-Hisp Races	1004	10%	951	10%	-5%
.	MISSING	359	4%	359	4%	---
		9658	**100%**	**9658**	**100%**	

The two distributions in Table 9d are not radically different, particularly for whites, blacks, and Asians, who comprise nearly 75% of the UW-BHS sample. For other groups, the effect of using coded write-ins, and thus the amount of measurement error in the unedited (raw) variables, is quite significant. The size of the NHOPI. AIAN, and Hispanic populations are all 10-12% larger when the content of write-ins is ignored. These net error rates understate the gross number of errors, moreover. Among Hispanics, e.g., more than a dozen respondents wrote a Hispanic identity strictly on the race question (under SOR). These Hispanics are not identified as such using the location-based measures, since the substance of the write-ins is not considered. Adding these Hispanics to the earlier total would slightly increase the number of Hispanics. But far more respondents made the opposite "error"—writing a non-Hispanic identity (like Italian) in the Hispanic Origin section. Subtracting these errant identifiers results in a net loss of more than 100 Hispanics, or 10% of the original (location-based) count.

The AIAN and NHOPI populations are also reduced by double digit percentages when coded write-ins are used to create **r15xgrpall—the cleaned summary race/ethnic classification based on content rather than locaiton**. In the latter case, a sizeable number of Indonesian respondents (who are coded as Asian in OMB guidelines) wrote their identity in the NHOPI write-in linethusinflating the NHOPI count in the location-based measure. Of course the largest changes are centered in the SOR "population." While nearly 2% of the UW-BHS sample used the "other race" write-in space as their sole means of identifying themselves, only 46 respondents failed to write something that could not be interpretable as (recoded to) an OMB category. Thus, large SOR count in the location-based measures is strictly an artifact of white, black, etc. respondents using unconventional means to express themselves. Still, it is questionable whether

these should be viewed as errors, strictly speaking, since many of the respondents who used the SOR space probably did so for a reason. Middle Eastern respondents, for example, may not see themselves as white, even if OMB considers that to be their "correct" racial identity.

As with the two sets of indicator variables from which they are derived, the choice of whether to use location- or content-based summary measures ultimately rests with the researcher. About 4% of UW-BHS respondents have a different identity in **r15xgrpall** than they have in **idin00815x.** The discrepancy between summary variables based on cleaned and unedited responses rises to more than 15% when additional write-ins, such as those on the ancestry and parental ancestry questions, are also taken into account, though most of this difference owes to the higher number of multiple ancestry persons identified in these extra measures. The summary variable that incorporates additional identities is called **r1xxgrpall. r1xxgrpall** is similar to **r15xgrpall** but uses the "1xx" (based on responses in additional questions 15x and 16x) indicators instead of the 15x indicators which are limited to race and Hispanic origin responses. The **r1xx…** summary measures are not covered in this memo (yet).

Table 9e: Variable r15xdet: Single OMB Race or Hispanic origin, and Major Combinations

Code	Group	N	Percent
101	White Alone	5152	53%
201	Black Alone	843	9%
301	AIAN Alone	65	1%
401	Asian Alone	1232	13%
501	NHOPI Alone	124	1%
601	SOR Alone	46	0%
701	Hisp Alone	381	4%
10100	Hisp/White	203	2%
10200	Hisp/Black	66	1%
10300	Hisp/AIAN	22	0%
10400	Hisp/Asian	39	0%
10999	Hisp/Other Combos	122	1%
11200	White/Black	203	2%
11230	White/Black/AIAN	65	1%
11240	White/Black/Asian	24	0%
11300	White/AIAN	200	2%
11400	White/Asian	262	3%
11450	White/Asian/NHPI	32	0%
11500	White/NHPI	27	0%
12300	Black/AIAN	47	0%
12400	Black/Asian	54	1%
14500	Asian/NHPI	30	0%
18060	Other Non-Hsp Combos	60	1%
.	MISSING	359	4%
		9658	**100%**

Table 9e shows the second content-based race/ethnicity variable likely to be of interest to researchers, **r15xdet** (**det**ailed), which retains the single race counts from **r15xgrpall** but breaks out Hispanic and (non-Hispanic) multiracial persons into detailed combinations. The figures show, for instance, that fewer than half of all Hispanics report exclusively Hispanic origins. Nearly one quarter identify as Hispanic and white, and the rest identify other combinations of Hispanic/racial origins. Non-Hispanic multiracials span a number of unique combinations, though nearly two in three are white/black, white/AIAN or white/Asian.

Table 9e also illustrates the coding scheme for multiracial (including race + Hispanic origin) persons in our race/ethnic taxonomy. The logic is as follows. First, all multi codes have five digits and begin with a 1 (as shown in Table 9c, the general "Multiracial" code is 10000). Second, the substance of the codes is the permutations of digits 1-6, which correspond exactly to the census categories white(1), black(2), AIAN(3), Asian(4), NHOPI(5) and SOR(6). These are the same root codes used in the three digit classifications discussed in section III (see Table 6a), which in turn were borrowed from census and IPUMS. In the multiracial codes, each component group is represented by a single digit, and as a general rule, the more non-zero digits, the more groups are in the combo. Thus:

1**12**00 = white/black
1**123**0 = white/black/AIAN
1**2345** = black/AIAN/Asian/NHOPI

The codes are arrayed hierarchically (privileging lower numbers over higher) and do not take ordering into account. This means that races with lower numbers "absorb" the combinations from races with higher ones. Thus, there are more combos grouped under white than under black, more under black than under Asian, etc. This is why white/black (11200) is coded but black/white is not. This is also why most Asian (code 4) and NHOPI (code 5) combos are grouped with the races whose codes precede them (e.g. white, black). The logic here was to subordinate SOR combos (code 6) to all others (hence no codes have a 6 that precedes one of the other major racial groups' numbers (1-5)).

Trailing zeros are meaningless in our coding scheme (e.g. the zeros in 11200 and 11300 are just place holders), but Hispanic combos use a leading "0" instead of the usual "7." Thus, Hispanic combos are of the form "10xxx" rather than “17xxx.” This change places Hispanic combos ahead of all other groups (including whites) when sorting the codes from lowest to highest. There are two reasons we prefer this ordering. The first is to keep Hispanic combos ahead of SOR combos (coded 6s) since the latter are not substantive identities. The second is to make it easy to strip off Hispanic combinations for users who prefer pooling all Hispanics regardless of race(s), leaving only non-Hispanic multiracials.

The content-based measures have strengths and limitations relative to the unedited race/ethnicity data. On the one hand, coding written identities significantly reduces the number of unusable (SOR) responses, and probably “corrects” a large number of measurement errors. These corrections rely on summary judgments of the UW-BHS staff, however. While we follow sensible reallocation rules used by the census and other agencies, there is always risk in editing the way individuals chose to identify themselves.

IV.C: Advanced OMB summary variables: Improving coverage and reducing ambiguity

Even after coding written identities, a number of problems remain with the basic content-based summary measures. In **r15xdet** and **r15xgrpall,** 359 cases have no valid race/Hispanicity data, and a small number (46) still have no OMB codable identity.

About 14% of the sample has multiple identities of some sort, either two or more OMB+ races or a combination of race and Hispanic origin. When Hispanics (of any race or races) are pooled, as in **r15xgrpall**, persons who identify any Hispanic roots are automatically coded strictly as Hispanic, rather than as members of other racial groups they may also identify.

Subsequent custom race measures attempt to resolve these ambiguities using various substitution and simplification rules. All of these summary measures begin with **r15xdet** (or **r15xgrpall**); they simply collapse combinations and/or substitute missing/SOR values with valid data from other sources. Two such summary variables deal only with combinations of standard OMB races: **r15x8privnw** and **r15x8privw**. These two variables employ opposite tie-breaking rules to reassign persons of white/non-white descent to a single race. **r15x8privnw priv**ileges the respondent's **n**on-**w**hite identity, while **r15x8privw priv**ileges **w**hite identity.

Table 9f shows the distribution of the variable **r15x8privnw**. Since the majority of multiracial persons have just two races, one of which is white, **r15x8privnw** dramatically reduces the number of multiracial respondents--from 10% of the UW-BHS sample to just over three percent.

Table 9f: r15x8privnw=r15x8grpall with White/Nonwhite multis assigned to Non-White

			Freq.	Percent	Valid	Cum.
Valid	101	White Alone	5152	53.34	55.40	55.40
	207	Black or Wht/Blk	1046	10.83	11.25	66.65
	307	AIAN or Wht/AIAN	265	2.74	2.85	69.50
	407	Asian or Wht/Asian	1494	15.47	16.07	85.57
	507	NHOPI or Wht/NHPI	151	1.56	1.62	87.19
	601	SOR Alone	46	0.48	0.49	87.69
	709	Hisp, Any Races	833	8.62	8.96	96.64
	18020	2+ Non-Wht Races	312	3.23	3.36	100.00
	Total		9299	96.28	100.00	
Missing	.		359	3.72		
Total			9658	100.00		

The coding rule in **r15x8privnw** is somewhat arbitrary. Treating all bi-racial persons as non-white by default is only reasonable if they see themselves strictly as race/ethnic minorities. Some may identify primarily as white. Thus, a more balanced approach is to take respondents' identity preferences directly into account. We accomplish this by using primary race/ethnicity as a tie breaker. This approach allows each respondent to simplify his/her identity on a case by case basis, rather than using a blanket transformation rule for all multiracial persons. Also, because the primary race/ethnicity question is asked of all respondents and is completely open ended, it can be used to simplify Hispanic/Non-Hispanic combinations, as well as combinations of standard OMB races (e.g. black & white).

The variable **r15x8prim** (translation: **r**ace from the **15x** questions, sorting respondents into **8** categories using **prim**ary identity to break ties) is shown in Table 9g. This variables codes persons with combination identities (including Hispanics) into the first OMB identity they report on the primary race/ethnicity question. For those who fail to supply a codable primary race, we use their first Ancestry response as a substitute. In all, 1352 multiracial (or mixed race/Hispanic) respondents are simplified using these two supplemental variables— 1158 via primary identify, and another 194 via first ancestry. As shown in the table, this transformation reduces the number of multiracial respondents by nearly

90%, from over 1000 cases to just 104. Notably, the number of Hispanics is also reduced significantly (from 833 to 539), since Hispanics who choose something else (e.g. white or black) as their primary identity are reassigned.

Table 9g: r15x8prim -- r15x8grpall with multi respondents coded to 1st OMB Primary(n=1158), or 1st Ancestry (n=194) if primary is missing.

			Freq.	Percent	Valid	Cum.
Valid	102	White Primary	5598	57.96	60.20	60.20
	202	Black Primary	1198	12.40	12.88	73.08
	302	AIAN Primary	136	1.41	1.46	74.55
	402	Asian Primary	1490	15.43	16.02	90.57
	502	NHOPI Primary	188	1.95	2.02	92.59
	601	SOR Alone	46	0.48	0.49	93.09
	702	Hisp Primary	539	5.58	5.80	98.88
	18040	2+ IDs, No Primary	104	1.08	1.12	100.00
	Total		9299	96.28	100.00	
Missing	.		359	3.72		
Total			9658	100.00		

r15x8prim is a highly useful summary measure. It is based on cleaned (coded) responses to the standard race and Hispanic origin queries. Supplemental measures (primary identity and ancestry) are used only to resolve ambiguous responses to the standard items. Using **r15x8prim,** more than 94% of UW-BHS respondents can be classified into one of 6 mutually exclusive OMB race/ethnic groups. Most remaining cases are due to non-response (3.72%), and only a handful (1.5%) are SOR or multiracial after primary identities are taken into account.

Table 9h: r15x7smax -- r15x8prim with Missing/SOR recoded to Primary (n=69), Ancestry (n=37), or Mom's Ancestry (n=7).

			Freq.	Percent	Valid	Cum.
Valid	100	WHITE	5674	58.75	60.57	60.57
	200	BLACK	1212	12.55	12.94	73.51
	300	AIAN	138	1.43	1.47	74.99
	400	ASIAN	1503	15.56	16.05	91.03
	500	NHOPI	189	1.96	2.02	93.05
	700	HISPANIC	547	5.66	5.84	98.89
	18040	2+ IDs, No Primary	104	1.08	1.11	100.00
	Total		9367	96.99	100.00	
Missing	.		291	3.01		
Total			9658	100.00		

Subsequent measures attempt to illuminate the remaining missing, multiple, or ambiguous cases. **r15x7smax** (translation: **r**ace from the **15x** questions, sorting respondents into **7** categories using **s**elf-reported data to **max**imize coverage), shown in Table 9h, employs a sequence of substitutions to maximizes the number cases with a valid race/ethnic identity. In constructing **r15x7smax,** respondents who skipped the race and Hispanic origin question, or only supplied an uncodable (SOR) identity on these items, are coded to their first primary identity, if available. Those without a primary identity are coded to their first ancestry, and those without either are coded to their mother’s first ancestry (no additional cases could be recovered using father’s ancestry). This iterative substitution eliminates all remaining SORs and about 70 missing cases. **r15x7smax** utilizes all of the available

race/ethnicity items on the baseline survey, and represents the **max**imum coverage we can obtain using **s**elf-reported measures of race/ethnic identity. Only 291 cases, about 3% of the sample, are unaccounted for by **r15x7smax.** There are also 104 multiracial cases that cannot be further simplified.

Table 9i: r15x6sadmax = r15x7smax with Missing and Multis recoded to their Admin Race from school records (n=363)

		Freq.	Percent	Valid	Cum.
Valid	100 WHITE	5850	60.57	60.77	60.77
	200 BLACK	1325	13.72	13.76	74.54
	300 AIAN	146	1.51	1.52	76.05
	400 ASIAN	1547	16.02	16.07	92.13
	500 NHOPI	189	1.96	1.96	94.09
	700 HISPANIC	569	5.89	5.91	100.00
	Total	9626	99.67	100.00	
Missing	.	32	0.33		
Total		9658	100.00		

Additional content-based summary variables draw upon supplemental UW-BHS project sources to obtain race/ethnic identities for cases that could not be identified (or simplified) using the senior survey. The most traction is gained by using the linked administrative records from respondents' schools. Table 9i details the variable **r15x6sadmax** (translation: **r**ace from the **15x** questions, sorting respondents into **6** categories using **s**elf-reported and **ad**ministrative data to **max**imize coverage). This variable is ideal for users to wish to conduct broad racial comparisons using the largest reasonable sample size. By incorporating the school data, every remaining combination identity is resolved, and all but 32 missing cases are eliminated. Using **r15x6sadmax,** 99.7% of the UW-BHS sample can now be classified into a single OMB race/ethnic category.

The remaining handful of cases can be allocated using additional sources such as yearbooks (external race) and an analysis of ethnic surnames, though we do not consider these measures in detail here.

IV.D: Sub-OMB summary variables: Accounting for national/ethnic origins

The summary variables presented thus far have been limited to OMB categories and combinations. Readers will recall that the UW-BHS questionnaire listed more detailed categories and contained several open ended spaces for respondents to report additional ethnic, tribal, and national origins. In this section, we outline a series of "ethnic" summary measures, and in the next section we integrate the ethnic/national origin identities with major OMB categories and combinations to create a series of hybrid summary measures.

The *racial* summary measures code permutations of the content-based OMB indicator variables to sort respondents into broad categories and combinations. The *ethnic* summary measures decompose each indicator variable into detailed national/ethnic origin groups. For example, while **idasasn** distinguishes those who identify as Asian from those who do not, **easn15xdet** (translation: **Det**ailed **Asi**a**n e**thnicities from section **15x**) decomposes the aggregate Asian population into specific national/regional origin groups (e.g. Chinese, Japanese, Multi-ethnic).

All ethnic variables have a complex but common construction. First, to ensure maximum coverage, we begin by coding a new set of OMB+ indicator variables that

codes the union of the location- and content-based indicators discussed in Section III. These new indicators are prefixed "**inas**" to designate that the respondent *either* identifies in the location for a particular group OR as a member of that group (regardless of location). For example, every respondent given a non-zero code for the Hispanic ethnicity variable **ehsp15xdet** made some claim (plausible or otherwise) of Hispanic origin, either by identifying as Hispanic or supplying any response other than "non-Hispanic" on the Hispanic origin question. Second, after using the **inas** indicators to identify the maximum potential count of each OMB race/pan-ethnic group, we analyze responses to the specific national/ethnic/tribal indicator variables derived from our race/ethnic taxonomy (see Table 8c). Recall that most of these indicators are derived from specific groups listed on the race and Hispanic origin questions (e.g. Chinese, Mexican, Samoan), while others are write-ins that occurred with sufficient frequency to merit their own code (e.g. Cherokee, Panamanian). Still others are custom codes denoting multiple specific origins (Chinese & Cambodian), or generic responses (writing "Hispanic" or marking "Other Asian" but writing nothing).

Finally, we create summary ethnicity variables for each of the six OMB+ categories. The codes used in these variables are based on the detailed sub-OMB identities claimed for each broad race/pan-ethnic category. Table 9j lists the ethnic summary variables and associated indicators derived from responses to the race and Hispanic origin questions only (15x). Corresponding measures based on the entire senior survey (including the 16x questions) are available but not discussed here.

Table 9j: Summary Ethnicity Variables and Source Variables: 15x info only

	Hispanic	AIAN	Asian	NHOPI	SOR
Grand Indicator Variable (1)	inashsp15x	inasami15x	inasasn15x	inasopi15x	inassor15x
Detailed Summary Ethnicity Variable (2)	ehsp15xdet		easn15xdet		esor15xdet
General Summary Ethnicity Variable	ehsp15xgen	eami15xgen	easn15xgen	eopi15xgen	
Dichotomous Ethnic Origin Indicators					
Specific Origins	idasmexi15x	idascher15x	idasasin15x	idasnhaw15x	idasuncd15x
	idasprcn15x	idasbkft15x	idascamb15x	idasgmch15x	idasdtno15x
	idascuba15x	idasotrb15x	idaschin15x	idassamo15x	idasumix15x
	idasscam15x	idasaknv15x	idasfili15x	idasospi15x	idasamer15x
	idaspana15x		idasjapa15x		idasjoke15x
	idasotsc15x		idaskore15x		idasaver15x
	idasohsp15x		idasloat15x		
	idasbzpt15x		idasviet15x		
			idasindo15x		
			idasthai15x		
			idasoasn15x		
Generic Origins	idasghsp15x	idasgami15x	idasgasn15x	idasgopi15x	
Multiple Origins	idasmhsp15x	idasmtrb15x	idasmasn15x	idasmopi15x	idasmsor15x

(1) Union of "idin" and "idas" indicators

(2) Detailed Summary variables have more categories than the general versions

Since respondents can fall under the heading of a broad pan-ethnic/racial group by identifying as a member of a group or in the corresponding location for that group,

it is possible for a person to have only one substantive OMB race/pan-ethnicity but several non-zero detailed ethnicity codes. For example, a respondent who checked white and wrote Irish under “Other Spanish…” AND “Other Asian” would only be coded as white in the content-based indicators (**idas…**) as well as the content-based summary race variables discussed in the previous subsection. But under the broader, more inclusive ethnicity variables, this respondent would receive a residual (non-zero) code under both the AIAN and Asian headings. The code labels for such respondents clarify our lack of faith in the validity of the claim, however: in each instance, a person whose tribe or Asian origin was “Irish” (or similar) would be coded “LE” (likely errant) for **eami15x** and **easn15x**.

Table 9k: Five Summary Ethnicity variables: 15x Section only

ehsp15xdet -- Detailed Hispanic Ethnicity, regardless of race, using info from 15x

			Freq.	Percent	Valid	Cum.
Valid	0	Non-Hisp	8343	86.38	89.72	89.72
	711	Mexican	426	4.41	4.58	94.30
	715	Puerto Rican	127	1.31	1.37	95.67
	718	Cuban	24	0.25	0.26	95.92
	721	Panamanian	21	0.22	0.23	96.15
	739	Other S/C American Nation	80	0.83	0.86	97.01
	740	Other Hispanic	15	0.16	0.16	97.17
	741	Multi-ethnic Hispanic	26	0.27	0.28	97.45
	790	Non-specific Hispanic	114	1.18	1.23	98.68
	800	Brazil/Portugal/Unknown	20	0.21	0.22	98.89
	999	LE:Non-Hisp	103	1.07	1.11	100.00
	Total		9299	96.28	100.00	
Missing	.		359	3.72		
Total			9658	100.00		

eami15xgen -- Tribal Response, regardless of race, using info from 15x

			Freq.	Percent	Valid	Cum.
Valid	0	Non-AIAN	8764	90.74	94.25	94.25
	311	Cherokee	87	0.90	0.94	95.18
	312	Blackfoot	29	0.30	0.31	95.49
	379	Other Tribes	131	1.36	1.41	96.90
	380	Alaska Native*	21	0.22	0.23	97.13
	389	Multiple Tribes	26	0.27	0.28	97.41
	390	No Tribe	206	2.13	2.22	99.62
	999	LE:Non-AIAN	35	0.36	0.38	100.00
	Total		9299	96.28	100.00	
Missing	.		359	3.72		
Total			9658	100.00		

eopi15xgen -- General Pac.Isl. Origin, regardless of race, using info from 15x

			Freq.	Percent	Valid	Cum.
Valid	0	Non-NHPI	9006	93.25	96.85	96.85
	510	Native Hawaiian	80	0.83	0.86	97.71
	521	Guamanian/Chamorro	40	0.41	0.43	98.14
	522	Samoan	109	1.13	1.17	99.31
	549	Other PIs	14	0.14	0.15	99.46
	561	Multi PI	17	0.18	0.18	99.65
	590	Non-specific PI	7	0.07	0.08	99.72
	999	LE:Non-NHPI	26	0.27	0.28	100.00

	Total	9299	96.28	100.00	
Missing	.	359	3.72		
Total		9658	100.00		

easn15xdet -- Detailed Asian Ethnicity, regardless of race, using info from 15x

		Freq.	Percent	Valid	Cum.
Valid	0 Non-Asian	7507	77.73	80.73	80.73
	411 Cambodian	199	2.06	2.14	82.87
	412 Chinese	91	0.94	0.98	83.85
	413 Filipino	306	3.17	3.29	87.14
	414 Japanese	132	1.37	1.42	88.56
	415 Korean	408	4.22	4.39	92.95
	416 Laotian	28	0.29	0.30	93.25
	417 Vietnamese	278	2.88	2.99	96.24
	418 Asian Indian	45	0.47	0.48	96.72
	419 Indonesian	28	0.29	0.30	97.02
	421 Thai	15	0.16	0.16	97.18
	459 Other Asians	12	0.12	0.13	97.31
	462 Chinese/Cambodian	36	0.37	0.39	97.70
	463 Chinese/Vietnamese	14	0.14	0.15	97.85
	464 Chinese/Other(s)	78	0.81	0.84	98.69
	465 Other Combos	75	0.78	0.81	99.49
	490 Non-specific Asian	12	0.12	0.13	99.62
	999 LE:Non-Asian	35	0.36	0.38	100.00
	Total	9299	96.28	100.00	
Missing	.	359	3.72		
Total		9658	100.00		

sor15xdet -- SOR Sub-groups using info from 15x

		Freq.	Percent	Valid	Cum.
Valid	0 Non-SOR	8550	88.53	91.95	91.95
	601 SOR, no WI	7	0.07	0.08	92.02
	610 Uncodable	10	0.10	0.11	92.13
	620 Don't Know	2	0.02	0.02	92.15
	630 Unknown Mixture	28	0.29	0.30	92.45
	641 American/Human	27	0.28	0.29	92.74
	642 Nonsensical	10	0.10	0.11	92.85
	643 Aversion	1	0.01	0.01	92.86
	699 2+ SOR	1	0.01	0.01	92.87
	700 LE:Spec.Hispan.	167	1.73	1.80	94.67
	701 LE:'Hispanic'	148	1.53	1.59	96.26
	800 LE:White Ethnic	170	1.76	1.83	98.09
	801 LE:'White'	21	0.22	0.23	98.31
	802 LE:NH 2+ Race	130	1.35	1.40	99.71
	999 LE:NH 1 Race	27	0.28	0.29	100.00
	Total	9299	96.28	100.00	
Missing	.	359	3.72		
Total		9658	100.00		

Table 9k lists frequency distributions for all five summary ethnicity variables. As discussed above, these variables assign codes to everyone who claims a valid identity within a given OMB group OR writes something in the space designed for

that group. Notably, only the former responses are given codes that correspond to an actual OMB category, however. Take the case of **eamigen15x**, the summary ethnicity variable for American Indians. This variable codes six subgroups that we consider to be "valid" AIAN identities, including tribes (Cherokee and Blackfoot), super-tribal groups (Alaska Natives and Multi-tribal Indians), and even non-specific "American Indians" (No Tribe). All of these codes are prefixed with the number 3, the root AIAN code in our race/ethnic taxonomy. Remaining respondents are given the residual (999) code "LE: Non-AIAN" to indicate that while they may have written something on the AIAN section of the survey, the substance of their write-ins leads us to believe that they are not, in fact, American Indian.

Other summary ethnicity variables are coded in a similar fashion. The rule is that persons coded "1" on the content-based indicator of a particular race, whom we believe to be actual members the group, are given 1) a substantive code indicating their specific national/tribal/ethnic origin or 2) a generic code indicating non-specific, but still plausible membership within the parent race category. Looking at **easn15xdet**, for example, we see that most respondents report one or more specific Asian national origins, and a dozen simply mark "other Asian" and move on. The sum of respondents with positive, non-"errant" codes on **easn15xdet** is identical to the sum coded 1 on **idasasn15x**, the content-based indicator of substantive Asian membership. But **easn15xdet** also illuminates persons who are Asian via location only, i.e. those for whom **idasasn15x = 0** but **idinasn = 1.** As shown in Table 9-l, there are 35 respondents who fit this profile, i.e. who write non-Asian origins beneath the "other Asian" line. These 35 respondents are listed only as "Likely Errors: Non-Asians" in **easn15xdet.** Note, however, that there are 33 additional "off diagonal" respondents—those who identified *as Asian* but not in the Asian section of the race question (e.g. by writing Thai under SOR or Indonesian under NHOPI). These respondents are coded alongside those who identified these groups in the more conventional manner: by marking the listed category or writing the identity under the appropriate (Asian, in this case) section of the race question.

Table 9-l: Cross-tabulation of Content- and Location- based Indicators of Asian origin

ID's as Asian on any s15x Section	**Supplies Any Response on Asian Section**			
	0	1	.	Total
0	7,507	**35**	0	7,542
1	**33**	1,724	0	1,757
.	0	0	359	359
Total	7,540	1,759	359	9,658

The coded SOR ethnicity variable, **esor15xdet,** warrants particular mention. The vast majority of SOR write-ins are, in fact, interpretable as OMB identities. This is why most responses in the SOR section are considered "likely errors" (in other words, not really "SOR"). The "true" SOR population, if it can even be described as such, is comprised of an eclectic collection of respondents who provide hard-to-code identities such as "American" or "unknown" when responding to race/ethnicity queries.

9.D: Hybrid Summary variables: Race & Ethnicity

The ethnicity variables provided detailed insights into the existence, number, and character of ethnic, national, and/or tribal identities that UW-BHS respondents

report while responding to standard questionnaire items on race and Hispanic origin. They distinguish specific ethnic responses from generic ones, and mono-ethnic origins from multi-ethnic roots. They even identify persons who are likely included under a given OMB+ category by mistake.

Users must bear in mind, however, that the ethnicity variables decompose *all* responses that fall under a given OMB identity or location, regardless of how many other OMB groups might have also been reported. The ethnicity variables are just that—a measure of the particulars of one's Asian, Hispanic, AIAN etc. identity, NOT the exclusivity of those identities with respect to other major race/pan-ethnic groups. Thus, a person coded as having Cherokee ethnicity on **eami15xgen** could have identified solely as Cherokee, or he could have marked white, black, Hispanic, and also wrote Cherokee.

As such, the OMB racial/pan-ethnic summary variables (shown in section 9.B) and the detailed ethnicity/nationality variables shown in section 9c depict opposite sides of a coin. The former tell us which race (or combination) a person identifies, regardless of ethnicity (e.g. Chinese, Japanese, Korean, all coded as "Asian"), while the latter tell us which ethnicity a person identifies, regardless of race (e.g. Black & Chinese, White & Chinese, or just Chinese, all coded as "Chinese"). Users who wish to account for the number, combination, and character of both OMB and sub-OMB identities either need to cross-tabulate the race and ethnicity variables, or construct "hybrid" variables of their own. In this section, we outline three such variables frequently used by UW-BHS staff: **eracedet**, **eracegen**, and **eracemax**.

Eracedet

The first is **eracedet** (translation: **eth**nicity and **race**, **d**etailed), shown in Table 9D-1. The construction of **eracedet** is complex, but the guiding principles are straightforward. At its core, **eracedet** is a cross-classification of race and ethnicity—specifically the cleaned OMB race variable **r15xdet** and the summary ethnicity variables **ehsp15xdet, eami15xgen, easn15xdet, eopi15xgen, esor15xdet**.

Eracedet attempts to preserve salient cultural and social distinctions within and between major OMB race/pan-ethnic groups without creating a variable that is too unwieldy or difficult to collapse. While **eracedet** contains 53 categories, this is a tremendous simplification of the actual permutation of the variables described above, which results in thousands of unique, but far too nuanced, identities. In constructing **eracedet**, our guiding logic was empirical as well as substantive. All groups with fewer than 20 observations are collapsed into a higher level of aggregation, and every group with 20 or more is given a distinctive code. Since the final configuration of categories and combinations is somewhat arbitrary, the numeric codes in **eracedet** do not correspond to the race/ethnic taxonomy used up to this point (though the root codes for the 7 OMB+ categories occupy the second digit of each value of **eracedet**).

Eracedet is a content-based measure derived from responses to the race and Hispanic origin sections (15x). It is constructed by decomposing the broad racial/pan-ethnic categories and combinations in **r15xdet** into detailed ethnic sub-groups. As such, observations are distinguished in *sequence*. The first distinction is between racial/pan-ethnic categories (including SOR) and combinations thereof. Any racial group or combination of sufficient size (n>19) is included in **eracedet**. Examples include Black/Asian and White/NHOPI. OMB Combinations that number fewer than 20 are pooled into one of two residual combination categories: those for Hispanic/race combinations, and those for two or more non-Hispanic race combinations.

Table 9D-1: eracedet – Race, Ethnicity, and Combinations: Detailed

		Freq.	Percent	Valid	Cum.
Valid	3100 White	5152	53.34	55.40	55.40
	3200 Black	843	8.73	9.07	64.47
	3202 Black+White	203	2.10	2.18	66.65
	3203 Black+AIAN	47	0.49	0.51	67.16
	3204 Black+Asian	54	0.56	0.58	67.74
	3205 Blk+Wht+Asn	24	0.25	0.26	68.00
	3300 AIAN:1+ Tribe	37	0.38	0.40	68.39
	3301 AIAN:No Tribe	28	0.29	0.30	68.70
	3302 I+W: Cherokee	42	0.43	0.45	69.15
	3303 I+W: OthTribe	91	0.94	0.98	70.13
	3304 I+W: No Tribe	67	0.69	0.72	70.85
	3305 I+W+B:1+ Tribe	29	0.30	0.31	71.16
	3306 I+W+B:No Tribe	36	0.37	0.39	71.55
	3400 Asn:Indian	37	0.38	0.40	71.94
	3401 Asn:Cambodian	193	2.00	2.08	74.02
	3402 Asn:Chinese	53	0.55	0.57	74.59
	3403 Asn:Filipino	182	1.88	1.96	76.55
	3404 Asn:Japanese	29	0.30	0.31	76.86
	3405 Asn:Korean	269	2.79	2.89	79.75
	3407 Asn:Vietnamese	269	2.79	2.89	82.64
	3408 Asn:Indonesian	20	0.21	0.22	82.86
	3409 Asn:China+Cambo	36	0.37	0.39	83.25
	3410 Asn:China+1+	57	0.59	0.61	83.86
	3411 Asn:2+ Ethnic	48	0.50	0.52	84.37
	3412 Asn:Other	39	0.40	0.42	84.79
	3413 A+W:Korean	87	0.90	0.94	85.73
	3414 A+W:Filipino	60	0.62	0.65	86.37
	3415 A+W:Japanese	57	0.59	0.61	86.99
	3416 A+W:Other(s)	58	0.60	0.62	87.61
	3500 NHPI:Guamanian	23	0.24	0.25	87.86
	3501 NHPI:Samoan	76	0.79	0.82	88.68
	3502 NHPI:Other(s)	25	0.26	0.27	88.95
	3503 PI+W	27	0.28	0.29	89.24
	3504 PI+A	30	0.31	0.32	89.56
	3505 PI+W+A	32	0.33	0.34	89.90
	3700 Hsp:Mexican	247	2.56	2.66	92.56
	3701 Hsp:P.Rican	38	0.39	0.41	92.97
	3702 Hsp:Lat.Amr	53	0.55	0.57	93.54
	3703 H+W:Mexican	90	0.93	0.97	94.50
	3704 H+W:P.Rican	26	0.27	0.28	94.78
	3705 H+W:Lat.Amr	26	0.27	0.28	95.06
	3706 H+B Mexican	22	0.23	0.24	95.30
	3707 H+B:Others	44	0.46	0.47	95.77
	3708 H+I	22	0.23	0.24	96.01
	3709 H+A	39	0.40	0.42	96.43
	3710 H+*:Mexican	42	0.43	0.45	96.88
	3711 H+*:P.Rican	34	0.35	0.37	97.25
	3712 Hsp:NonSpec	25	0.26	0.27	97.52
	3713 H+W:NonSpec	44	0.46	0.47	97.99
	3714 H+*:NonSpec	23	0.24	0.25	98.24
	3799 H|H+*:Other(s)	58	0.60	0.62	98.86
	3800 SOR Alone	46	0.48	0.49	99.35
	3801 2+ NH Races	60	0.62	0.65	100.00
	Total	9299	96.28	100.00	
Missing	.	359	3.72		
Total		9658	100.00		

The second dimension by which respondents are distinguished is ethnicity. Within each major OMB+ category and combination, every ethnic, national, or tribal sub-group of 20 or more is broken out and given a detailed sub-code. Thus, code 3403 “Asn:Filipino” denotes respondents who marked/wrote Asian alone and identified Filipino as their sole ethnic identity, while code 3409 (Asn:China+Cambo) denotes persons whose only major race is still Asian, but who have two Asian ethnicities (Chinese and Cambodian). Other codes signal multiple OMB identities in combination with one or more detailed ethnic identities. For example, code 3704 “H+W:P.Rican” denotes respondents who identified as white and Hispanic and marked/wrote Puerto Rican as their specific Hispanic ethnicity, while code 3305 “I+W+B:1+ Tribe” denotes respondents who marked white, black, and AIAN and then wrote one or more specific tribes.

Since group size is a guiding principle in choosing which categories are broken out in **eracedet**, the resulting array of identities is not always intuitive. It is, nonetheless, *informative*. For example, there is no code for monoracial Indians who identify as Cherokee, though there is code for White/Indian Cherokees. The reason, of course, is that most Cherokees are persons of multiracial descent. In fact, no Indian tribe has even 20 monoracial persons in the UW-BHS data, and the sum of all single race persons with a tribal identity is only 37. Even some listed categories failed to register 20 or more monoracial, mono-ethnic respondents; hence Cubans and Laotians are pooled with other small and/or intermixed groups.

Other points to bear in mind while interpreting **eracedet**:

1) Broad Racial/pan-ethnic identities are to the left of the colon, while detailed ethnic/national identities are to the right.
2) All groups are monoracial and/or mono-ethnic unless denoted with a “+” sign (or other indicator)
3) W=white, B=black, I=Am. Indian, A=Asian, PI=NHOPI, H=Hispanic
4) * = “all other races” + = AND | = OR 1+ = One or more

Thus, Asn:China+1+ is interpreted as Single Race Asians with Chinese plus one or more additional Asian origins. H|H+*:Other(s) denotes persons who are Hispanic alone or Hispanic in combination with other OMB groups, and who have a Hispanic ethnicity other than those already listed.

Eracedet illuminates the racial and ethnic distinctions (and intersections) of UW-BHS respondents in a highly detailed manner. Researchers interested in honing in on various axes of differentiation will likely find **eracedet** quite useful. The dimensions embedded within the variable already reveal interesting dynamics between the context and character of race/ethnic identities. For example, we see that monoracial Hispanics (those who identify soley as Hispanic) almost always identity a specific Hispanic origin, while multiracial Hispanics (those who also identify as white or some other race(s)) are much more likely to claim a vague/generic Hispanic identity. These seemingly subtle variations in identity may reflect meaningful differences in migration experience, interracial parentage, and/or the strength of community attachments between monoracial and multiracial Hispanics. **Eraced** can be used to draw insights into these and other questions. For example, does language usage and school performance vary between monoracial and mixed race Hispanics? Do American Indians who maintain a tribal affiliation differ from those who do not?

Eracegen

Not all researchers need to distinguish race/ethnic populations at the level of detail provided by **erecedet**, however. The variable contains dozens of racial, ethnic, and race/ethnic combinations, which adds a great deal of complexity to most

analyses. When so much variation *within* major race/pan-ethnic categories, efforts to summarize differences between major subpopulations are rendered more difficult.

Thus, we have also constructed two additional, simpler hybrid race/ethnicity variables: **eracegen** and **eracemax**. Both variables are similar, though the latter augments the former by using school records to resolve ambiguous/missing cases.

Eracegen is conceptually similar to **eracedet** in that it is also derived from a cross-tabulation of race and ethnicity. However, while **eracedet** starts with all of the single race categories and combinations (**r15xdet**), **eracegen** is derived from the much simpler variable **r15x7smax** (Table 9h), which reassigns most of the missing, multiple, and ambiguous identities into a single (or primary) OMB+ category. Since **r15x7smax** has far fewer to categories to begin with, its decomposition into ethnic subgroups results in a much more parsimonious race/ethnicity variable.

Table 9-D2: General Race/Ethnicity based on r15x7smax & e15x ethnicity variables

			Freq.	Percent	Valid	Cum.
Valid	100	WHITE	5674	58.75	60.57	60.57
	200	BLACK	1212	12.55	12.94	73.51
	310	American Indian Tribe	85	0.88	0.91	74.42
	390	Non-specific American Indian	53	0.55	0.57	74.99
	411	Cambodian	269	2.79	2.87	77.86
	412	Chinese	85	0.88	0.91	78.77
	413	Filipino	269	2.79	2.87	81.64
	414	Japanese	88	0.91	0.94	82.58
	415	Korean	370	3.83	3.95	86.53
	416	Laotian	22	0.23	0.23	86.76
	417	Vietnamese	283	2.93	3.02	89.78
	418	Asian Indian	39	0.40	0.42	90.20
	419	Indonesian	26	0.27	0.28	90.48
	421	Thai	16	0.17	0.17	90.65
	459	Other/Multi/Generic Asian	36	0.37	0.38	91.03
	510	Native Hawaiian	38	0.39	0.41	91.44
	521	Guamanian/Chamorro	30	0.31	0.32	91.76
	522	Samoan	91	0.94	0.97	92.73
	549	Other/Multi/Generic PI	30	0.31	0.32	93.05
	711	Mexican	339	3.51	3.62	96.67
	715	Puerto Rican	71	0.74	0.76	97.43
	721	Panamanian	17	0.18	0.18	97.61
	739	Other Latin American	75	0.78	0.80	98.41
	740	Other/General Hispanic	45	0.47	0.48	98.89
	18040	2+ IDs, No Primary	104	1.08	1.11	100.00
	Total		9367	96.99	100.00	
Missing	.		291	3.01		
Total			9658	100.00		

Eracegen is constructed as follows: we start with the six OMB+ categories defined by **r15x7smax**. Recall that this variable pools persons who identify a major race alone and those who identify it in combination but still choose it as their primary identity. Thus WHITE is taken to mean those who are only white as well as those who are primarily white. Next, within each single/primary race defined by **r15x7smax**, each ethnic/tribal/national origin with 15 persons or more is broken out into the detailed code specified by the corresponding **exxx15x** ethnicity variables (**ehsp15x**, **emai15x**, **easn15x**, & **eopi15x**). Thus, single/primary Asian (from **r15x7smax)** is decomposed into single/primary Chinese, Korean, or Japanese, etc. The final step is to resolve multi-ethnic (but same OMB) identities. Many respondents are monoracial at the OMB level but claim multiple ethnic origin groups. These “multi-ethnic” identities, shown in variables such **easn15xdet** and **ehsp15xdet** can only be resolved

by looking at the detailed, coded responses to the open ended questions on primary identity and ancestry, which we use as tie breakers.[16] Thus a respondent who marked Chinese and Cambodian but listed Chinese as their primary identity would be simplified as Chinese.

The end result, shown in Table 9D-2, is a variable with fewer than half the number of categories as **eracedet. Eracegen** contains no SORs and very few multiracial and/or multi-ethnic persons. Furthermore, by drawing on additional survey questions (not just the 15x section), only a small number of respondents from each major OMB+ group are left with a residual ("other…")code. Relative to **eracedet,** counts of every major race and sub-OMB ethnic/national group increase considerably, since we are now taking respondent's preferences for their "primary" identity into account. All told, less than 5% of the UW-BHS population cannot be coded into a single race/ethnic group: 104 multiracials who refused to provide a primary (or first ancestry, etc.), and 322 respondents who

Eracemax

The final hybrid race/ethnicity variable resolves all but a couple dozen of these remaining cases. Starting with **eracegen**, **eracemax** substitutes missing and remaining (unreducible) multiracial identities with the single race/ethnicity code from respondents school records. After this adjustment, only 28 cases remain uncoded.

eracemax -- eracegen with missing/multis assigned to school admin race

		Freq.	Percent	Valid	Cum.
Valid	100 WHITE	5850	60.57	60.75	60.75
	200 BLACK	1326	13.73	13.77	74.52
	310 American Indian Tribe	85	0.88	0.88	75.40
	390 Non-specific American Indian	63	0.65	0.65	76.05
	411 Cambodian	269	2.79	2.79	78.85
	412 Chinese	85	0.88	0.88	79.73
	413 Filipino	269	2.79	2.79	82.52
	414 Japanese	88	0.91	0.91	83.44
	415 Korean	370	3.83	3.84	87.28
	416 Laotian	22	0.23	0.23	87.51
	417 Vietnamese	283	2.93	2.94	90.45
	418 Asian Indian	39	0.40	0.40	90.85
	419 Indonesian	26	0.27	0.27	91.12
	421 Thai	16	0.17	0.17	91.29
	459 Other/Multi/Generic Asian	80	0.83	0.83	92.12
	510 Native Hawaiian	38	0.39	0.39	92.51
	521 Guamanian/Chamorro	30	0.31	0.31	92.82
	522 Samoan	91	0.94	0.94	93.77
	549 Other/Multi/Generic PI	31	0.32	0.32	94.09
	711 Mexican	339	3.51	3.52	97.61
	715 Puerto Rican	71	0.74	0.74	98.35
	721 Panamanian	17	0.18	0.18	98.53
	739 Other Latin American	75	0.78	0.78	99.30
	740 Other/General Hispanic	67	0.69	0.70	100.00
	Total	9630	99.71	100.00	
Missing	.	28	0.29		
Total		9658	100.00		

[16] The substitution procedure is as follows: if one primary is written, we code the multiracial/multi-ethnic respondent to that primary identity. If two primaries are written, we code the one that comes first. If no primaries are written, we code the first ancestry. If none of the above are written, we code mother's first ancestry, or father's.

APPENDIX

Partial Custom Race Variable Listing

Name	Label
Location-based Indicator Variables	
idinwht	Supplied Response on s159a:White
idinblk	Supplied Response on s159b:Black
idinhsp	Supplied Response on s158a-b:Hisp
idinami	Supplied Response on s159c-d:AIAN
idinasn	Supplied Response on s159e-n:Asian
idinopi	Supplied Response on s159o-s:NHOPI
idinsor	Supplied Response on s159u-t:SOR
idin159	Supplies Response on Race Question
idin15x	Supplies Response on R&H Section
idin16x	Supplies Response on any s16x Section
idin1xx	Supplies Response on any s15x or s16x Section
idinref	Supplies Response on s162: Reflected Identity
idinadm	R&E data from Adminstrative Records
idinoms	R&E data from OMS
Location Summary Race Variables	
idin008159	8-cat Summary of s159 ID Locations (as-is, no cleaning)
idin00815x	8-cat Summary of s158-s159 ID Locations (as-is, no cleaning)
idin009159	RECODE of idin008159
Content Indicator Variables	
idaswht15x	ID's as White on s15x Section
idaswht1xx	ID's as White on s15x or s16x Section
idasblk15x	ID's as Black on s15x Section
idasblk1xx	ID's as Black on s15x or s16x Section
idasami15x	ID's as AIAN on s15x Section
idasami1xx	ID's as AIAN on s15x or s16x Section
idasasn15x	ID's as Asian on s15x Section
idasasn1xx	ID's as Asian on s15x or s16x Section
idasopi15x	ID's as NHPI on s15x Section
idasopi1xx	ID's as NHPI on s15x or s16x Section
idassor15x	ID's as SOR on s15x Section
idassor1xx	ID's as SOR on s15x or s16x Section
idashsp15x	ID's as Hispanic on s15x Section
idashsp1xx	ID's as Hispanic on s15x or s16x Section
idasmexi158	Identifies as Mexican in s158

idasprcn158	Identifies as Puerto Rican in s158			
idascuba158	Identifies as Cuban in s158			
idasscam158	Identifies with a S/C American Nation in s158			
idaspana158	Identifies as Panamanian in s158			
idasotsc158	Identifies with another S/C Amer. Nation in s158			
idasohsp158	Identifies with another Hispanic Origin in s158			
idasghsp158	Identifies simply as 'Hispanic' in s158			
idasbzpt158	Identifies as Brazilian or Portuguese in s158			
idasmhsp158	Identifies multiple Hispanic Origins in s158			
idasmexi15x	Identifies as Mexican in s15x			
idasprcn15x	Identifies as Puerto Rican in s15x			
idascuba15x	Identifies as Cuban in s15x			
idasscam15x	Identifies with a S/C American Nation in s15x			
idaspana15x	Identifies as Panamanian in s15x			
idasotsc15x	Identifies with another S/C Amer. Nation in s15x			
idasohsp15x	Identifies with another Hispanic Origin in s15x			
idasghsp15x	Identifies simply as 'Hispanic' in s15x			
idasbzpt15x	Identifies as Brazilian or Portuguese in s15x			
idasmhsp15x	Identifies multiple Hispanic Origins in s15x			
idasmexi1xx	Identifies as Mexican in s1xx			
idasprcn1xx	Identifies as Puerto Rican in s1xx			
idascuba1xx	Identifies as Cuban in s1xx			
idasscam1xx	Identifies with a S/C American Nation in s1xx			
idaspana1xx	Identifies as Panamanian in s1xx			
idasotsc1xx	Identifies with another S/C Amer. Nation in s1xx			
idasohsp1xx	Identifies with another Hispanic Origin in s1xx			
idasghsp1xx	Identifies simply as 'Hispanic' in s1xx			
idasbzpt1xx	Identifies as Brazilian or Portuguese in s1xx			
idasmhsp1xx	Identifies multiple Hispanic Origins in s1xx			
idasmtrb15x	Identifies multiple Tribes in s15x			
idasmtrb1xx	Identifies multiple Tribes in s1xx			
idascher15x	Identifies as Cherokee in s15x			
idasbkft15x	Identifies as Blackfoot in s15x			
idasotrb15x	Identifies some other Tribe in s15x			
idasaknv15x	Identifies as Alaska Native in s15x			
idasgami15x	Identifies simply as 'Indian' or 'Native Amer.' in s15x			
idascher1xx	Identifies as Cherokee in s1xx			
idasbkft1xx	Identifies as Blackfoot in s1xx			
idasotrb1xx	Identifies some other Tribe in s1xx			
idasaknv1xx	Identifies as Alaska Native in s1xx			
idasgami1xx	Identifies simply as 'Indian' or 'Native Amer.' in s1xx			

idasasin15x	IDs as Asian Indian on s15x				
idascamb15x	IDs as Cambodian on s15x				
idaschin15x	IDs as Chinese on s15x				
idasfili15x	IDs as Filipino on s15x				
idasjapa15x	IDs as Japanese on s15x				
idaskore15x	IDs as Korean on s15x				
idasloat15x	IDs as Laotion on s15x				
idasviet15x	IDs as Vietnamese on s15x				
idasindo15x	IDs as Indonesian on s15x				
idasthai15x	IDs as Thai on s15x				
idasoasn15x	IDs some other Asian Origin on s15x				
idasgasn15x	IDs simply as 'Asian' on s15x				
idasmasn15x	IDs Multiple Asian Origins on s15x				
idasasin1xx	IDs as Asian Indian on s1xx				
idascamb1xx	IDs as Cambodian on s1xx				
idaschin1xx	IDs as Chinese on s1xx				
idasfili1xx	IDs as Filipino on s1xx				
idasjapa1xx	IDs as Japanese on s1xx				
idaskore1xx	IDs as Korean on s1xx				
idasloat1xx	IDs as Laotion on s1xx				
idasviet1xx	IDs as Vietnamese on s1xx				
idasindo1xx	IDs as Indonesian on s1xx				
idasthai1xx	IDs as Thai on s1xx				
idasoasn1xx	IDs some other Asian Origin on s1xx				
idasgasn1xx	IDs simply as 'Asian' on s1xx				
idasmasn1xx	IDs Multiple Asian Origins on s1xx				
idasnhaw15x	IDs as Nat. Hawaiian on s15x				
idasgmch15x	IDs as Guamanian on s15x				
idassamo15x	IDs as Samoan on s15x				
idasospi15x	IDs Other PI Origin on s15x				
idasgopi15x	IDs simply as 'NHOPI' on s15x				
idasmopi15x	IDs Multiple PI Origin on s15x				
idasnhaw1xx	IDs as Nat. Hawaiian on s1xx				
idasgmch1xx	IDs as Guamanian on s1xx				
idassamo1xx	IDs as Samoan on s1xx				
idasospi1xx	IDs Other PI Origin on s1xx				
idasgopi1xx	IDs simply as 'NHOPI' on s1xx				
idasmopi1xx	IDs Multiple PI Origin on s1xx				
idasmsor15x	Supplies Multiple SOR responses in s15x				
idasmsor1xx	Supplies Multiple SOR responses in s1xx				
idasuncd15x	Supplies Uncodable response (usually religion) in s15x				

idasdtno15x		Writes 'Don't Know' or 'Unknown' in s15x						
idasumix15x		Writes 'Mixed' or similar in s15x						
idasamer15x		Writes 'American' or 'Human' in s15x						
idasjoke15x		Supplies Non-sensical response (Martian, elf, e.g.) in s15x						
idasaver15x		Expresses Aversion toward R/E questions in s15x						
idasuncd1xx		Supplies Uncodable response (usually religion) in s1xx						
idasdtno1xx		Writes 'Don't Know' or 'Unknown' in s1xx						
idasumix1xx		Writes 'Mixed' or similar in s1xx						
idasamer1xx		Writes 'American' or 'Human' in s1xx						
idasjoke1xx		Supplies Non-sensical response (Martian, elf, e.g.) in s1xx						
idasaver1xx		Expresses Aversion toward R/E questions in s1xx						
Content-based Summary Race Variables								
r15xdet		Single OMB Race/Hisp|No OMB Race/Hisp|OMB Combinations @s158-s159						
r15x8grpall		Non-Hisp Single Race|Any Hispanic|Pooled Multirace @s15x						
r15x8privnw		r15x8grpall w/MultiRule:White+Nonwhite=>Non-White						
r15x7bhpia		r15x8grpall w/Rule:Black>Hisp>NHPI>AIAN>Asian						
r15x8privwh		r15x8grpall w/MultiRule:Wht+Nonwht => White						
r15x8prim		r15x8grpall w/Rule:1st OMB Primary(n=1158)>>1st Ancestry(n=194)						
r15x7prhr		r15x8prim>>r15x7bhpia(n=104) if primary/ancestry missing						
r15x7smax		r15x8prim w/Miss|SOR:Primary(n=69)>Ancestry(n=37)>Mom's Anc(n=7)						
r15x6sadmax		r15x7smax w/Miss|Multis w/o Primary => Admin Race(n=363)						
r15x6nomiss		r15x6sadmax with Missing=OMS(n=24)>Name(n=7)>White(n=2)						
r15x7bhpiapct~s		Percent of OMS Ratings that match Self-reported Race in r15x7bhpia						
r15x7prhrpctobs		Percent of OMS Ratings that match Self-reported Race in r15x7prhr						
r15x7smaxpctobs		Percent of OMS Ratings that match Self-reported Race in r15x7smax						
r15x6sadmaxpc~s		Percent of OMS Ratings that match Self-reported Race in r15x6sadmax						
r15x6nomisspc~s		Percent of OMS Ratings that match Self-reported Race in r15x6nomiss						
r1xxdet		Single OMB Race/Hisp|No OMB Race/Hisp|OMB Combinations @s15x-16x						
r1xx8grpall		Non-Hisp Single Race|Any Hispanic|Pooled Multirace @s15x-16x						
r1xx8privnw		r1xx8grpall w/MultiRule:White+Nonwhite=>Non-White						
r1xx7bhpia		r1xx8grpall w/Rule:Black>Hisp>NHPI>AIAN>Asian						
r1xx8privwh		r1xx8grpall w/MultiRule:Wht+Nonwht => White						
r1xx8prim		r1xx8grpall w/Rule:1st OMB Primary(n=2172)>>1st Ancestry(n=351)						
Content-based Summary Ethnicity Variables								
ehsp15xdet		Detailed Hispanic Ethnicity, regardless of race, using info from 15x						
ehsp15xgen		General Hispanic Ethnicity, regardless of race, using info from 15x						
eami15xgen		Tribal Response, regardless of race, using info from 15x						

easn15xgen	General Asian Ethnicity, regardless of race, using info from 15x
easn15xdet	Detailed Asian Ethnicity, regardless of race, using info from 15x
eopi15xgen	General Pac.Isl. Origin, regardless of race, using info from 15x
esor15xdet	SOR Sub-groups using info from 15x
Content-based Summary Race/Ethnicity Variables	
eracedet	Detailed Race/Ethnicity based on r15xdet & e15x ethnicity variables
eracegen	General Race/Ethnicity based on r15x7smax & e15x ethnicity variables
eracemax	eracegen with missing/multis assigned to school admin race