Identifying Race and Ethnicity in the 1979 National Longitudinal Survey of Youth

Audrey Light
Alita Nandi

Department of Economics and Center for Human Resource Research
The Ohio State University

February 2004

Address correspondence to Audrey Light, Department of Economics, Ohio State University, 1945 N. High Street, Columbus, OH 43201 or light.20@osu.edu. We thank Rosella Gardecki, Steve McClaskie and Karima Nagi for valuable input.

Abstract

The 1979 National Longitudinal Survey of Youth is among the few surveys to provide multiple reports on respondents' race and ethnicity. Respondents were initially classified as Hispanic, black, or "other" on the basis of data collected during 1978 screener interviews. Respondents subsequently self-reported their "origin or descent" in 1979, and their race and Hispanic origin in 2002; the latter questions conform to the federal standards adopted in 1997 and used in the 2000 census. We uses these data to (a) assess the size and nature of the multiracial population, (b) measure the degree of consistency among these alternative race-related variables, and (c) devise a number of alternative race/ethnicity taxonomies and determine which does the best job of explaining variation in log-wages. A key finding is that the explanatory power of race and ethnicity variables improves considerably when we cross-classify respondents by race *and* Hispanic origin. Little information is lost when multiracial respondents are assigned to one of their reported race categories because they make up only1.3% of the sample.

I. Introduction

Over the last 25 years, federal standards for measuring race and ethnicity in the United States underwent two major changes. In 1977, the Office of Management and Budget mandated that individuals should be classified *separately* by race and ethnicity (OMB 1977). The standards at that time required the use of at least four race categories (white, black, American Indian or Alaska native, and Asian or Pacific Islander) and two ethnicity categories (Hispanic origin or not of Hispanic origin). As a result, individuals could be cross-classified as white and Hispanic or black and non-Hispanic, for example, but not as white and black. Twenty years later, in what has been termed "the greatest change in the measurement of race in the history of the United States" (Farley 2002), the standards were revised to allow individuals to report more than one race (OMB 1997).[1] The new standards apply to administrative agencies throughout the country as well as to virtually all federally-funded surveys that collect race data, including the decennial Census of Population, the Current Population Surveys, and the National Longitudinal Surveys.

These sweeping changes led to renewed interest in the analysis of race and ethnicity. The majority of recent studies ask what the "new" data on race and ethnicity reveal about the U.S. population. Studies that describe the multiracial population and characterize the racial distribution of Hispanics include Farley (2002), Goldstein and Morning (2000), and Waldrop and Long (2002). In addition, an extensive literature has emerged on bridging methods that reassign multiple-race respondents to a single race category (Allen and Turner 2001; Grieco 2002; Lee 2001; OMB 2000; Tucker *et al*. 2002). These methods simplify the race taxonomy by eliminating multiple-race categories, and allow researchers to maintain a uniform race distribution across data collection regimes. Because the emphasis has been on population-wide assessments, recent race-related studies focus almost exclusively on data from the 2000 census.[2]

In this study, we extend the analysis of race and ethnicity to the 1979 National Longitudinal Survey of Youth (NLSY79). In contrast to the decennial census and other cross-sectional surveys, the NLSY79—which has followed several thousand individuals from 1979 to the present—provides multiple reports on each sample member's race and ethnicity. Sample members were initially classified as Hispanic, black, or non-Hispanic/nonblack on the basis of

[1]Additional changes made in 1997 include separating Asian and Pacific Islander (termed "Native Hawaiian or other Pacific Islander") into at least two categories, and renaming the ethnicity categories "Hispanic or Latino" and "not Hispanic or Latino."

[2]Race/ethnicity studies that do *not* use census data include Scott (1999) and Telles and Lim (1998).

information collected during 1978 screener interviews. During the 1979 interviews, respondents reported their "origin or descent." Twenty-three years later, in 2002, respondents answered questions on Hispanic origin and race that conform to the 1997 federal standards. This sequence of variables provides a unique opportunity to learn how respondents' racial and ethnic classifications change over time and in response to different questions.

Our analysis of the NLSY79 race and ethnicity data proceeds in three stages. Following recent census-based analyses, we first describe the multiracial population revealed by the 2002 self-reports, and determine how the race distribution is affected by alternative bridging methods. Second, we assess the degree of internal consistency in the 1978, 1979, and 2002 race-related variables. We determine how often an individual's racial or ethnic classification varies across these alternative reports, and whether the inconsistencies are more common among identifiable groups such as Hispanics, American Indians, and/or multiracial individuals. Third, we devise a number of alternative racial/ethnic taxonomies that bring to bear the patterns revealed in steps 1 and 2, and ask which does the best job of explaining variation in log-wages—an outcome that is among the most frequently studied by social scientists, often with a focus on racial and ethnic disparities (see, for example, Altonji and Blank 1999; Heckman *et al*. 2000; Smith and Welch 1989; Trejo 1997). We consider very simple classifications (*e.g.,* Hispanic vs. non-Hispanic and white vs. nonwhite) as well as detailed schemes in which individuals are cross-classified according to a single race (white, black, Asian, *etc.*) and Hispanic origin. We then consider even finer classification schemes that distinguish between single-race and multiple-race individuals, and between individuals who are consistently coded across years and those for whom race is identified inconsistently. Our goal is to determine which classification scheme is "best" in the sense of maximizing the between-category variance (R^2) of our chosen outcome.

Our analysis has direct value to NLSY79 users, but we believe it is useful to *all* researchers using race and ethnicity data collected under the new federal standards. Even in the absence of longitudinal data where cross-year inconsistencies must be reconciled, researchers using "new" data invariably find that respondents report a staggering number of race-ethnicity combinations.[3] While other analysts have asked what these detailed data reveal about the U.S. population, we

[3]When five races are used and respondents are allowed to select between one and five races, a maximum of 31 single- and multiple-race categories can be formed. When cross-classified with two ethnic categories, this yields a 62-category race/ethnicity taxonomy. The taxonomy grows to 126 categories when a sixth race code (*e.g.*, "other" or "refuse") is added.

ask how analysts can wrestle this information into a manageable taxonomy for use in standard regression analysis. We believe our study is the first to provide guidance on maximizing the explanatory power of the racial/ethnic information collected under the new federal standards.

II. Data

Sample

The NLSY79 began in 1979 with a sample of 12,686 respondents who were born between 1957 and 1964. The original sample consists of three groups: a representative subsample of the civilian population in the designated birth cohort, an over-sample of 5,295 Hispanics, blacks, and economically disadvantaged non-Hispanics and nonblacks, and a subsample of 1,280 individuals who served in the military. Respondents were interviewed annually from 1979 to 1994 and biennially thereafter. Additional information about the survey can be found in Center for Human Resource Research (2001).

We confine our analysis to a subsample of 7,662 respondents who satisfy two criteria. First, they must be interviewed in 2002 (the twentieth interview round) because in that year respondents were asked two race/ethnicity questions that conform to the new federal standards. This eliminates 4,962 original sample members. The disadvantaged non-Hispanic, nonblack respondents and most of the military subsample were dropped in earlier rounds, and another 2,240 respondents left the survey of their own accord in 2002 or earlier. Second, a response must be coded for the race questions asked in 2002 and 1979. This rule excludes only 62 individuals, and allows us to maintain a uniform sample while comparing any pair of variables.

Race and ethnicity variables

In 2002, NLSY79 respondents were asked two race-related questions patterned after the race questions introduced in the 2000 census. These questions, along with the variable names that we assign them, are:

HISP02: *Are you Hispanic, Latino, or of Spanish origin?*

RACE02: *What race or races do you consider yourself to be?*

Respondents gave a simple yes/no response to the first question. This coding scheme differs from the 2000 census, in which respondents answered "yes" by selecting one of four Hispanic origins (Puerto Rican; Mexican, Mexican American or Chicano; Cuban; other). The second question (RACE02) was asked in an open-ended fashion—that is, respondents were neither shown a hand card nor read a list of options. Interviewers coded each answer into one of seven

categories: White; black or African American; Asian; native Hawaiian or other Pacific Islander; American Indian or Alaska Native; other; or "refuses to classify race."[4] In contrast, the 2000 census provided respondents with a list of 15 categories. The census has four options (native Hawaiian; Guamanian or Chamorro; Samoan; or other Pacific Islander) rather than a single "native Hawaiian or Pacific Islander" category, and in place of "Asian" it specifies Asian Indian, Chinese, Filipino, Japanese, Korean, Vietnamese, and other Asian as seven separate options (U.S. Bureau of the Census 2000). Throughout our analysis, we combine Asian and "native Hawaiian or other Pacific Islander" into a single group.

Prior to the 2002 interview, NLSY79 respondents were asked race-related questions only once, in 1979. The two questions asked in 1979 are:

RACE79: *What is your origin or descent?*

PRACE79: *You said that your origin or descent was ——. Which one of these do you feel closest to?*

The first question was asked of all respondents. They were shown a hand card with 30 categories (listed in table 1) and asked to select "all that apply." Respondents who chose more than one category were asked the second question, which is intended to identify their primary or preferred origin. To facilitate comparison with the 2002 responses, we aggregate the 30 categories coded in 1979 into the same seven categories (including other and none/refuse) used for RACE02. Table 1 indicates how we form these aggregates.

In addition to the four self-reported race variables described above, the NLSY79 provides an additional race variable that plays a central role in our analysis. This variable is:

RACE78: A *created* variable based on race and ethnicity information elicited during the 1978 screener; classifies sample members as Hispanic, black, or other (non-Hispanic/ nonblack).

This variable is created from four data items obtained during screener interviews conducted in 1978. First, respondents were asked to select their "origin or descent" from a list of 15 categories shown on a hand card. If a respondent (by which we mean a youth who was subsequently selected for the NLSY79 sample) was not present during the screener, an adult (usually the youth's parent) answered this question on his or her behalf. Second, the adult was asked whether he/she or his/her spouse spoke Spanish as a child. Third, the respondent's

[4]Interviewers were instructed to read these categories to the respondent (excluding "refuse") if the respondent did not provide an answer or the interviewer was unsure how to code the response.

surname was noted. Fourth, the interviewer recorded the respondent's race as white, black, or other based on inspection; if the respondent was not present, he/she was assigned the adult's race. Based on this information, respondents were deemed to be Hispanic if they had a Spanish surname, the adult respondent spoke Spanish, or they chose a Hispanic category as their origin or descent. Remaining respondents were coded as black if they chose "black, Negro, or Afro-American" as their origin or descent, or were identified by the interviewer as black. All other respondents were classified as non-Hispanic, nonblack.[5]

The variable that we refer to as RACE78 is the "official" race designation for NLSY79 respondents. It is used to compute sampling weights and to define the racial subsamples described in the preceding subsection. As a result, researchers have historically relied on this variable to assess racial and ethnic differences among NLSY79 respondents. Prior to the release of the 2002 data, variables available for identifying race/ethnicity were limited to RACE78, RACE79, PRACE79, and interviewer reports (recorded at every interview except 1987, 2000, and 2002) of each respondent's race as white, black, or other. Aside from RACE78, however, these variables are rarely used. In fact, RACE78 appears to be the *only* variable used to identify race and ethnicity of NLSY79 respondents among the race-related studies cited in Bureau of Labor Statistics (2002).

III. Summary of Race and Ethnicity Variables

We begin by summarizing the 2002 self-reports of race and Hispanic origin (RACE02 and HISP02). We then bring the remaining variables (RACE79, PRACE79, and RACE78) into the analysis to determine the extent to which each respondent's race and Hispanic origin is consistently identified.

2002 self reports

Table 2 summarizes responses to the race question (RACE02) asked in 2002—the only time NLSY79 respondents were asked to report their *race* rather than their "origin or descent." The columns in table 2 labeled "no assignment" indicate that 98.7% of sample members chose a single race category, while the remaining 1.3% chose two, three, or four categories. In contrast, 2.4% of respondents in the 2000 census chose two or more race categories, and six was the

[5] See Center for Human Resource Research (2001) for additional details on the screener interviews and the creation of RACE78. The variable RACE78 is referred to in the NLSY79 database and documentation as "R's racial/ethnic cohort from screener" (R02147), which is collapsed from the "sample identification code" (R01736).

maximum number of categories chosen (Grieco and Cassidy 2001). Although these census figures provide a useful benchmark, we do not expect the NLSY79 and 2000 census race distributions to match because, as noted in section II, the NLSY79 sample was (a) drawn more than 20 years before the 2000 census, (b) confined to an eight-year birth cohort, and (c) designed to over-sample blacks and Hispanics.[6]

Table 3 shows the race combinations chosen by the 101 respondents who identify themselves as multiracial. The four most frequently chosen combinations, in descending order, are white and American Indian, white and other, black and American Indian, and white and black; these four categories account for 73% of multiracial respondents. In the 2000 census, white and other, white and American Indian, white and Asian, and white and black are the four most common combinations, accounting for 72.4% of the multiracial population (Grieco and Cassidy 2001). Overall, 68% of multiracial respondents in the NLSY79 select white as one of their races, 44% select black, and 59% select American Indian.

Returning to table 2, we report the distributions that result from reassigning the multiracial respondents to appropriate one-race categories. We use a subset of the "bridging" methods applied to 2002 census data by Grieco (2002), Tucker *et al*. (2002), and others. First, we assign respondents to each of their chosen race categories—for example, we place the 26 respondents choosing both white and American Indian into *both* of those categories. This strategy, which we term maximum assignment, double-counts the 90 respondents who choose two race categories, and triple- or quadruple-counts the remaining 11 multiracial respondents. Next, we assign the multiracial respondents to the single category among their selected categories that has the fewest sample members (minority assignment) and, alternatively, to the category with the most sample members (majority assignment). Finally, we use a fractional assignment scheme in which, for example, respondents who select white and American Indian contribute half an observation to each category. Because only 1.3% of respondents are reassigned, the resulting race distributions are not very sensitive to the alternative assignment strategies—the percentage of respondents in each race category changes by less than one percentage point as we alter the assignment method. However, these reassignments represent a significant change among the race categories with

[6]A result of the NLSY79 sampling design is that whites and Asians are underrepresented. In contrast to the distribution in table 2, the breakdown among 2000 census respondents choosing a single race is 75.1% white, 12.3% black, 3.7% Asian/Hawaiian/Pacific Islander, and 0.9% American Indian, and 5.5% other (Grieco and Cassidy 2001).

very few respondents. By switching from majority assignment to maximum or minority assignment, we more than double the number of respondents considered to be American Indian and increase the count of Asian respondents by one third.

While RACE02 identifies only 101 multiracial respondents, table 2 reveals that almost 800 respondents (10.3% of the sample) report their race as "other" or simply refuse to report a race during the 2002 interview. A cross-tabulation of RACE02 and HISP02 (table 4) shows that the majority of these individuals are Hispanic. All but two of the 471 respondents who refuse to provide an answer to RACE02 identify themselves as Hispanic, while 86% of those selecting "other" as their race are Hispanic. In all, 93% of respondents who are coded as "refuse" or "other" in RACE02 (including those who report another race in addition to "other") self-report as Hispanic. Although the two questions were designed to cross-classify all respondents according to Hispanic origin *and* race, it is apparent that many Hispanic respondents do not identify as white, black, Asian/Pacific Islander, or American Indian; they view Hispanic as their origin *and* their race. This tendency has also been documented by Guzmán and McConnell (2002), Martin *et al.* (1990) and Rodriguez (1991).

Consistency of the 1978, 1979 and 2002 variables

We now ask whether the race information reported in 2002 is consistent with the information obtained earlier in the survey. To begin, we focus on the three indicators of Hispanic origin: the 2002 Hispanic origin question (HISP02), selection in 1979 of a Hispanic "origin or descent" (RACE79), and classification as Hispanic by RACE78. In table 5, we report the percent of respondents within each 2002 race category for whom all three indicators agree (*i.e.*, all three indicators classify the respondent as either Hispanic or non-Hispanic). Table 5 indicates that Hispanic origin is consistently coded for 96.7% of the entire sample. The rate of agreement is highest (98.9%) among blacks and lowest among Asians/Pacific Islanders and multiracial respondents.[7] Among the 252 cases (3.3% of the sample) where disagreement occurs, HISP02 is the variable most likely to be at odds with the other two. This is unsurprising, given that HISP02 was reported more than 20 years after the other information on Hispanic origin. It is worth noting that HISP02 is equally likely to be the only "yes" and the only "no" among the three indicators. In other words, there is no evidence of a trend toward respondents

[7]There are 13 cases of "disagreement" among the 134 respondents classified as Asian/Pacific Islander or multiracial. In 11 of these 13 cases, the respondent is identified as non-Hispanic by two of the ethnicity indicators.

"discovering" or "escaping" their Hispanic origins over time. More generally, table 5 suggests there is little ambiguity about which respondents are Hispanic, regardless of when or how the information is reported.

Turning from Hispanic origin to race, we first compare the 2002 self-reports (RACE02) to the 1979 self-reports (RACE79 and PRACE79). We expect these variables to be less consistent than the Hispanic origin indicators because of the nature of the race questions. Respondents were asked to report their "origin or descent" in 1979 and their race in 2002. In 1979 they were given a hand card that included several Hispanic categories (table 1), whereas in 2002 they were expected to report their race(s) *in addition* to their Hispanic origin. Moreover, respondents could select multiple categories in both 1979 and 2002; as a result, we must distinguish between cases where the races are identical and cases where the selected categories merely overlap.

In table 6, we report the percent of respondents who consistently report their race(s) in 1979 and 2002. We begin with a subsample of 6,178 respondents for whom this comparison is straightforward because they select a single category in both years.[8] Blacks have an extremely high rate of cross-year consistency: 97% of respondents who choose black as their (only) race in 2002 also select black as their (only) "origin or descent" in 1979. Among respondents who select white as their only race in 2002, less than two-thirds select "white" (English, German, *etc*.) as their origin in 1979. Instead, 16% report "other" or "none," 15% choose Hispanic, and 3% choose American Indian. Respondents who select Asian or American Indian in 2002 also have a relatively low rate of cross-year consistency; most of the discrepancies arise because Hispanic is selected in 1979. Only 3.9% of respondents who choose "other" as their race in 2002 report "other" in 1979, and only one respondent is coded as "refuse" (or "none") in both years. As we saw in table 4, most respondents for whom RACE02 is coded "other" or "refuse" are Hispanic—unsurprisingly, the majority (92%) of these respondents select Hispanic as their origin in 1979. This confounding of race and ethnicity accounts for most of the discrepancy between RACE02 and RACE79. Among the 4,442 respondents who choose white, black, Asian, or American Indian as their only race/origin in *both* years, the consistency rate is 96%.

The remaining columns of table 6 refer to the 1,484 respondents who select two or more categories for RACE02 and/or RACE79. For this subsample, we use two alternative definitions

[8]Respondents may select multiple categories in 1979 as long as their responses fall into a single *aggregate* category, as defined in table 1. If a respondent chooses English, German, and Irish as his "origin or descent," for example, we classify him as white only.

of consistency. First, we require each race reported in 2002 to be among the categories coded for RACE79. If a respondent reports white as his only race in 2002, for example, his 1979 report is deemed a match if he selects at least one "white" category (English, French, German, *etc*.), regardless of what other categories he selects. If a respondent reports both white and black in 2002, he must also choose white and black in 1979. Clearly, respondents who select multiple races in 2002 are relatively unlikely to report their races consistently over time under this definition. While 92% of the 2002 "single-race" respondents have consistent reports (meaning their 1979 races include the one selected in 2002), only 24% of the two-race respondents and none of the 3- or 4-race respondents are deemed to be consistent. In the right-most column of table 6, we consider an alternative definition of cross-year consistency: the *primary* race indicated by PRACE79 must be among the races reported in 2002. Individuals who choose a single race in 2002 have a lower rate of agreement under this definition, while those choosing multiple categories in 2002 have a higher rate. As more races are reported for RACE02, it becomes increasingly likely that one of those races is the primary race reported in 1979.

In table 7, we extend the comparison of race identifiers to include RACE78. Because RACE78 classifies respondents first as Hispanic and then as black or other, we can only ask whether *race* is coded consistently across the alternative variables (RACE78, RACE79 and RACE02) for non-Hispanics. Among the 4,953 non-Hispanics (as classified by HISP02) who report a single race in 1979 and 2002, the overall consistency rate is 85%. This is considerably higher than the 69% rate seen in table 6 because we eliminate most of the cases where respondents report "other" (for example) in 2002 but Hispanic in 1979.[9] Again, black respondents have near-perfect agreement among the multiple race indicators. Respondents reporting white, Asian, American Indian or other in 2002 must select the same category in 1979 and be coded as "other" by RACE78. For all categories except Asian, the consistency rates are higher than what is seen in table 6. Although there are only a handful of Asians in this subsample, a surprisingly high percentage of them are classified as Hispanic *or* black by RACE78. This inconsistent classification also holds for the few Asians in the subsample of 1,300 multiracial, non-Hispanics. In general, however, the three race indicators prove to be fairly consistent once we reduce the "noise" caused by Hispanic respondents.

[9]The table 7 subsample consists of respondents identified by HISP02 as non-Hispanic. Given the high rate of agreement among the Hispanic indicators, RACE78 and RACE79 are rarely coded as Hispanic.

IV. Explanatory Power of Alternative Race-Ethnicity Classification Schemes

The variables summarized in the preceding section allow us to identify respondents' race and ethnicity in numerous ways. Using HISP02, we can classify respondents as Hispanic or non-Hispanic. We can place each respondent into one of the six or seven single-race categories available for RACE02, using alternative bridging methods for the multiracial subsample; this race taxonomy can be collapsed into fewer categories, or expanded to include multiple-race categories. Alternatively, we can cross-classify respondents by race *and* ethnicity. For any taxonomy, we can reclassify respondents after switching to another variable—for example, we can use RACE78 or RACE79, rather than HISP02, to determine who is Hispanic. Rather than relying on a single variable, we can include separate categories for respondents who are identified inconsistently.

This range of possibilities exists because the NLSY79—and any survey that conforms to the new federal standards—provides extremely detailed data on race and ethnicity. While the detail is ideal for researchers wishing to study the multiracial population or describe the racial composition of Hispanics, it poses a challenge for researchers who simply require a manageable set of race/ethnicity variables for modeling a particular outcome. In this section, we ask how alternative race/ethnicity taxonomies differ in their ability to explain variation in one of the most widely studied outcomes in social science research: log-wages.

To carry out this exercise, we work with a reduced sample of 6,994 respondents who report an earned wage during their 2000 or 2002 interview. Our dependent variable is the natural logarithm of the most recently reported average hourly wage divided by the year-specific implicit price deflator for gross domestic product. We do not control for any factors other than race and ethnicity, but to reduce the effect of extreme outliers we top- and bottom-code wages at the values corresponding to the 98^{th} and 2^{nd} percentiles for our sample. In table 8, we report the portion of the variance in log-wages that is explained by alternative sets of race/ethnicity variables. The best taxonomy is the one with relatively little within-group variation in wages and, therefore, a relatively high portion of total variation that is "explained" by between-group differences. While explanatory power can invariably be improved by adding categories, our goal is to find a parsimonious way to capture most of the explanatory power contained in the most detailed classification scheme.

In the first row of table 8, we report the R^2 obtained by classifying respondents as Hispanic,

black, or "other" using RACE78. As noted in section II, this taxonomy has been used in virtually all NLSY79-based research to date. In subsequent rows of table 8, we report the R^2 associated with alternative classification schemes and the percent change relative to this benchmark value of 3.816.[10]

The next several rows in table 8 reveal that we lose considerable explanatory power upon switching to a two-way race taxonomy (either black vs. nonblack or white vs. nonwhite), and especially when we simply distinguish between Hispanics and non-Hispanics. These specifications (2a-c, 3a-c, and 4b-c) also reveal that, for each taxonomy, the R^2 is significantly higher when we use RACE02 or HISP02 to classify respondents rather than RACE78 or RACE79. This is true for all taxonomies that we have tried: groups are somewhat more homogenous (*i.e.,* within-group variation in log-wages is reduced) when we use the 2002 race and ethnicity data. As a result, we use RACE02 and HISP02 for each remaining specification described in table 8 with the exception of the bottom two rows, where we account for cross-year inconsistency in the race reports.

In rows 5 through 8b we classify respondents by race only, using increasingly detailed classification schemes. The three-way classification (white, black, and other) in row 5 performs considerably better than any of the two-way taxonomies, but the R^2 does not increase significantly relative to row 1 until we decompose the nonblack/nonwhite group into finer categories. In row 6a, we use five categories (Asian, American Indian, other, refuse, and an aggregation of all multiple-race combinations) in addition to white and black. In row 7a, we use the minority assignment method described in section III to classify all multiple-race respondents into a one-race category. In row 8a, we include all 12 multiple-race categories listed in table 3. While the R^2 increases as we move from 6a to 7a to 8a, the marginal gain associated with the inclusion of all 12 multiple-race categories does not justify the associated decrease in degrees of freedom. Specification 7a has almost as much explanatory power as 8a, yet it is the most parsimonious of these three. (Taxonomies based on other "bridging" methods yield similar results.) Moreover, we find that very little explanatory power is lost when we combine "other"

[10]It is unsurprising that less than 4% of the total log-wage variance is explained by these three race/ethnicity variables, given that richly-specified wage models typically produce a relatively modest R^2 of about 0.20-0.30 when micro-data of this nature are used. Using NLSY79 data, Light and Strayer (2004) find that a detailed schooling taxonomy explains 10% of the total variance in log-wages, while the addition of a host of other regressors raises the R^2 to 0.24.

and "refuse" into a single category (6b, 7b, and 8b); this is unsurprising, given that both groups consist primarily of Hispanics.

Before considering the remaining rows of table 8, we turn to table 9 for insights into how to increase explanatory power further. In table 9, we report the weighted, within-group variance in log-wages as a percent of total variance, where the weights are the fraction of the sample in the particular group; we present this decomposition for a subset of taxonomies only. Focusing on specification 5, we find that almost 60% of the total log-wage variance is due to variation within the "white" group, while only 25% is due to variation among blacks and 11% is due to variation among "others." Whites account for most of the unexplained variation because their log-wages have more variation around the group mean than do the other groups' log-wages, *and* because they have a larger sample share.[11] Clearly, we can realize the largest gains in R^2 by decomposing whites (as opposed to blacks or "others") into a more homogenous sub-group.

Cross-classifying by race and Hispanic origin is the most straightforward way to divide whites into smaller groups. Returning to table 8, we augment race-only specifications 5, 6b, 7b, and 8b by adding Hispanic interactions. In the specifications labeled with a "single prime" (5', 6b', *etc*.) we interact all race variables with a Hispanic origin indicator; in those labeled with a double prime (7b" and 8b") we only subset whites and the other/refuse group into Hispanics and non-Hispanics. Explanatory power improves considerably when respondents are cross-classified by race *and* ethnicity. Each R^2 is 4-5% higher than what we obtain with the corresponding "race only" specification, and with the exception of specification 5' each R^2 exceeds our benchmark by 9-12%.[12] Because 7b" uses only seven race/ethnicity categories to achieve an R^2 of 4.18, it appears to be a particularly useful taxonomy.

In the bottom two rows of table 8, we augment specifications 7b and 7b" by dividing whites into "consistent" and "inconsistent" subgroups, where "consistent" means white is the only race coded for both RACE79 and RACE02. We focus on whites because this group has a relatively low consistency rate (table 6) and accounts for most of the unexplained variance in log-wages (table 9). Accounting for cross-year consistency in this manner increases the R^2 by 5.8% relative

[11]The unweighted, within-group variance (as a percent of total variance) for whites is 102.5 (59.8/.584), which means this group's log-wage variance is 2.5% larger than the total log-wage variance.

[12]Comparing specifications 5 and 5' or 7b and 7b" in table 9, we see that the improvement is almost entirely due to dividing "whites" into two groups, although non-Hispanic whites still account for more than half of the unexplained variance in log-wages.

to 7b and 2.6% relative to 7b". When compared to the benchmark, the gains are 11.1% and 12.3%. The sample shares in table 9 reveal that we are essentially defining three types of whites: Hispanics, "consistent" non-Hispanics, and "inconsistent" non-Hispanics. The latter group, which accounts for about 15% of the "inconsistent" subsample, is made up of individuals who selected white as their only race in 2002, but reported other races (perhaps in combination with white) in 1979—in other words, they may be multiracial individuals who did not identify themselves as such in 2002. Based on these and other experiments, we believe it is advantageous to weigh the responses to both RACE79 and RACE02 before classifying respondents by race. A trivial number of blacks, Asians, and American Indians will be reclassified on the basis of 1979 reports, but the white, "other," and multiracial respondents can often be more cleanly identified by considering both variables.

V. Conclusions

We have conducted a detailed assessment of the race and ethnicity variables available in the NLSY79. We conclude by restating our key findings and offering advice to researchers working with race and ethnicity data collected under the new federal standards.

- Very few individuals self-identify as multiracial: 2.4% of individuals in the 2000 census select more than one race, and only 1.3% of NLSY79 respondents do so. In all likelihood, researchers wishing to study the multiracial population will have to use census data to obtain a large enough sample. For many research purposes, however, multiracial respondents can simply be assigned to one of their reported race categories. Reassignment has virtually no effect on the race distribution or on the explanatory power of race variables in a log-wage model.

- One goal of the new federal standards is to classify individual by race and ethnicity, but a significant number of individuals decline to report a race. Race is coded "other" or "refuse" for 10.3% of NLSY79 respondents. Among respondents who refuse to report a race or who select "other" (with our without another race category), 93% are Hispanic. This indicates that the multiracial population is even smaller than it first appears, for many individuals who select "other" along with a race (*e.g.*, white) are, in fact, Hispanic and monoracial.

- These reporting problems notwithstanding, researchers who include race/ethnicity variables in standard regression models are advised to cross-classify respondents by race and Hispanic origin. In our log-wage models, we achieve the largest gains in R^2 when we interact

race indicators with Hispanic origin indicators; these taxonomies perform significantly better than one that simply controls for Hispanic, black, and other. It is particularly important to subset whites into Hispanic and non-Hispanics because this group accounts for a much larger share of total log-wage variance than does any other race group.

- We conclude that race and ethnicity data are not very sensitive to how and when the information is obtained (although we note an important exception below). The NLSY79 collected three Hispanic origin indicators over 24 years, yet 97% of respondents are coded consistently as Hispanic or non-Hispanic by all three measures. Respondents self-reported their "origin or descent" in 1979 and their race in 2002. Among those choosing a single, "official" race (white, black, Asian/Pacific Islander, or American Indian) in both years, the consistency rate among the two self-reports is 96%. However, the consistency rates fall dramatically for multiracial individuals and Hispanics—who, as discussed above, often select "other" or refuse to report their race if they are unable to report it as Hispanic. In assessing the explanatory power of alternative race/ethnicity taxonomies, we realized substantial gains by using the multiple reports to distinguish between "consistently reported" whites and those who may instead be multiracial. (We also find that that taxonomies based on the 2002 race/ethnicity data perform better than those based on data collected at the outset of the survey.)

Although we hope our study is useful to researchers working with race and ethnicity data from surveys other than the NLS79, the extent to which the patterns seen here apply to other data sources is unknown. As other longitudinal surveys provide multiple reports on respondents' race and ethnicity, we can learn more about the consistency of the data over time and across data collection regimes. In addition, we can assess the explanatory power of alternative race/ethnicity taxonomies using data from the 2000 census and other surveys, and using alternative outcome measures. These extensions will enhance our understanding of how individuals perceive their racial and ethnic identities, and how researchers can use the available data as productively as possible.

References

Allen, James P. and Eugene Turner. 2001. "Bridging 1990 and 2000 Census Race Data: Fractional Assignment of multiracial populations." *Population and Research Policy Review* 20(6): 513-533.

Altonji, Joseph G. and Rebecca M. Blank. 1999. "Race and Gender in the Labor Market." Pp. 3143-3259 in *Handbook of Labor Economics*, Volume 3C, edited by Orley C. Ashenfelter and David Card. New York: Elsevier Science Press.

Bureau of Labor Statistics. 2002. *The NLS Annotated Bibliography*. Washington, D.C.: Bureau of Labor Statistics, U.S. Department of Labor. Available online at http://www.nlsbibliography.org /index.php3.

Center for Human Resource Research. 2001. *NLSY79 User's Guid*e. Columbus, OH: Ohio State University. Available online at http://www.bls.gov/nls/79guide/2001/nls79g0.pdf.

Farley, Reynolds. 2002. "Racial Identities in 2000: The Response to the Multiple-Race Response Option." Pp. 33-61 in *The New Race Question: How the Census Counts Multiracial Individuals*, edited by Joel Perlmann and Mary C. Waters. New York: Russell Sage Foundation.

Goldstein, Joshua R. and Ann J. Morning. 2000. "The Multiple-Race Population of the United States: Issues and Estimates." *Proceedings of the National Academy of Sciences of the United States of America* 97(11): 6230-6235.

Grieco, Elizabeth M. 2002. "An Evaluation of Bridging Methods Using Race Data from Census 2000." *Population and Research Policy Review* 20(1-2): 91-107.

Grieco, Elizabeth M. and Rachel C. Cassidy. 2001. "Overview of Race and Hispanic Origin." Census 2000 Brief. U.S. Census Bureau. Available online at http://www.census.gov/prod/2001pubs/c2kbr01-1.pdf.

Guzmán, Betsy and Eileen Diaz McConnell. 2002. "The Hispanic Population: 1990-2000 Growth and Change." *Population Research and Policy Review* 21(1-2): 109-128.

Heckman, James J., Thomas M. Lyons and Petra E. Todd. 2000. "Understanding Black-White Wage Differentials: 1960-1990." American Economic Review 90(2): 344-349.

Lee, Sharon M. 2001. *Using the New Racial Categories in the 2000 Census.* Washington, D.C.: Annie E. Casey Foundation and Population Reference Bureau.

Light, Audrey and Wayne Strayer. 2004 "Who Receives the College Wage Premium? Assessing the Labor Market Returns to Degrees and College Transfer Patterns." *Journal of Human Resources* 39(3).

Martin, Elizabeth, Theresa J. DeMaio and Pamela C. Campanelli. 1990. "Context Effects for Census Measures of Race and Hispanic Origin." *Public Opinion Quarterly* 54(4): 551-566.

Office of Management and Budget (OMB). 1977. "Directive Number 15, Race and Ethnic Standards for Federal Statistics and Administrative Reporting." Washington, D.C.: U.S. Office of Management and Budget.

Office of Management and Budget (OMB). 1997. "Revisions to the Standards for the Classification of Federal Data on Race and Ethnicity." Washington, D.C.: U.S. Office of Management and Budget.

Office of Management and Budget (OMB). 2000. "Provisional Guidance on the Implementation of the 1997 Standards for Federal Data on Race and Ethnicity." Washington, D.C.: U.S. Office of Management and Budget.

Rodriguez, Clara E. 1991. "Race, Culture, and Latino 'Otherness' in the 1980 Census." *Social Science Quarterly* 73(4): 930-937.

Scott, Charles G. 1999. "Identifying the Race or Ethnicity of SSI Recipients." *Social Security Bulletin* 62(4): 9-20.

Smith, James P. and Finis Welch. 1989. "Black Economic Progress After Myrdal." *Journal of Economic Literature* 27(2): 519-564.

Telles, Edward E. and Nelson Lim. 1998. "Doest it Matter Who Answers the Race Question? Racial Classification and Income Inequality in Brazil." *Demography* 35(4): 465-474.

Trejo, Stephen J. 1997. "Why Do Mexican Americans Earn Low Wages?" *Journal of Political Economy* 105(6): 1235-1268.

Tucker, Clyde, Steve Miller and Jennifer Parker. 2002. "Comparing Census Race Data Under the Old and New Standards." Pp. 365-390 in *The New Race Question: How the Census Counts Multiracial Individuals*, edited by Joel Perlmann and Mary C. Waters. New York: Russell Sage Foundation.

U.S. Bureau of the Census. 2000. "United States Census 2000 Short Form Questionnaire." Available online at http://www.census.gov/dmd/www/pdf/d61a.pdf.

Waldrop, Judith and John F. Long. 2002. "A First Look at the 21st Century: Census 2000." *Population Research and Policy Review* 21(1-2): 3-16.

Table 1: Aggregate Race Categories Based on 1979 Self-Reported "Origin or Descent"

Race category (RACE79)	Origin or descent listed on hand card
White	English; French; German; Greek; Irish; Italian; Polish; Portuguese; Russian; Scottish; Welsh.
Black	Black, Afro-American, or Negro
Asian/Pacific Islander	Chinese; Filipino; Hawaiian or Pacific Islander; Indian-Asian; Japanese; Korean; Vietnamese.
American Indian	Indian-American or Native American
Hispanic	Chicano; Cuban or Cubano; Mexican or Mexicano; Mexican-American; Puerto Rican; Other Latino, Hispano or Latin American; Other Spanish descent.
Other	Other (specify); (If volunteered) American
None	None

Table 2: Race Distribution Based on 2002 Self-Reports
Using Alternative Assignment of Multiple-Category Responses

2002 self-reported race (RACE02)	No assignment		Maximum assignment		Minority assignment		Majority assignment		Fractional assignment	
	Num.	Pct.	Num.	Pct.	Num.	Pct.	Num.	Pct.	Num.[a]	Pct.
1-race categories										
White	4411	57.6	4480	57.6	4411	57.6	4480	58.5	4444	58.0
Black	2285	29.8	2329	30.0	2293	29.9	2314	30.2	2305	30.1
Asian/Pacific Islander	33	0.4	44	0.6	44	0.6	33	0.4	38	0.5
American Indian	49	0.6	109	1.4	108	1.4	52	0.7	77	1.0
Other	312	4.1	342	4.4	335	4.4	312	4.1	326	4.3
Refuse to report	471	6.2	471	6.1	471	6.2	471	6.2	471	6.2
All 1-race categories	7561	98.7	7775	100.0	7662	100.0	7662	100.0	7662	100.0
All 2-race categories	90	1.2								
All 3- & 4-race categories	11	.1								
Total	7662	100.0								

Note: Respondents selecting multiple races are alternately assigned to each reported category (maximum assignment), the category with the fewest respondents (minority assignment), the category with the most respondents (majority assignment) and each reported category with a weight of 1/n, where n is the number of races reported (fractional assignment).

[a]Rounded to the nearest whole number.

Table 3: Distribution of Multiple-Category Race Responses for 2002 Self-Reports

2002 self-reported race (RACE02)	Num.	Pct.
2-race categories		
White + Black	8	7.9
White + Asian/Pacific Islander	6	5.9
White + American Indian	26	25.7
White + Other	20	19.8
Black + Asian/Pacific Islander	4	4.0
Black + American Indian	20	19.8
Black + Other	3	3.0
American Indian + Other	3	3.0
3- & 4-race categories		
White + Black + American Indian	6	5.9
White + American Indian + Other	2	2.0
Black + American Indian + Other	2	2.0
White + Black + Asian/PI + Amer. Indian	1	1.0
All multiple race categories	101	100.0

Table 4: Percent Hispanic Based on 2002 Self-Reports by 2002 Self-Reported Race

2002 self-reported race (RACE02)	Hispanic (HISP02)	
	Num.	Percent
1-race categories		
White	4411	13.8
Black	2285	1.1
Asian/Pacific Islander	33	15.2
American Indian	49	20.4
Other	312	85.9
Refuse to report	471	99.6
2-race categories		
White + Black	8	0.0
White + Asian/Pacific Islander	6	0.0
White + American Indian	26	15.4
White + Other	20	65.0
Black + Asian/Pacific Islander	4	0.0
Black + American Indian	20	0.0
Black + Other	3	0.0
American Indian + Other	3	100.0
3- and 4-race categories		
White + Black + American Indian	6	16.7
White + American Indian + Other	2	50.0
Black + American Indian + Other	2	0.0
White + Black + Asian/PI + Amer. Indian	1	100.0
Total reporting other/refuse	813	92.7
Total	7662	18.4

Table 5: Percent of Respondents for Whom Hispanic Indicators Are in Agreement, by 2002 Self-Reported Race

2002 self-reported race (RACE02)	Num.	Percent Agree
1-race categories		
White	4411	96.1
Black	2285	98.9
Asian/Pacific Islander	33	87.9
American Indian	49	98.0
Other	312	92.6
Refuse to report	471	95.8
All 1-race categories	7561	96.8
All 2-race categories	90	92.2
All 3- and 4-race categories	11	81.8
Total	7662	96.7

Note: "Agree" means the respondent is identified by all three indicators (RACE78, RACE79 and HISP02) as either Hispanic or non-Hispanic.

Table 6: Comparison of 1979 and 2002 Self-Reported Race

2002 self-reported race (RACE02)	Report 1 race in 1979 and 2002		Report >1 race in 1979 and/or 2002		
	Num.	2002 race identical to 1979 race	Num.	2002 races among 1979 races[a]	2002 races include 1979 primary race[b]
1-race categories					
White	3214	65.7	1197	97.0	49.5
Black	2194	96.6	91	93.4	70.3
Asian/Pacific Islander	24	79.2	9	77.8	33.3
American Indian	30	56.7	19	94.7	68.4
Other	280	3.9	32	18.8	0.0
Refuse	436	0.2	35	0.0	0.0
All 1-race categories	6178	69.3	1383	92.3	48.7
All 2-race categories			90	24.4	67.8
All 3-, 4-race categories			11	0.0	90.9
Total	6178	69.3	1484	87.5	50.1

[a]All 2002 races are reported in 1979; may have 1979 races not reported in 2002.
[b]A race reported in 2002 is reported as the "primary race" in 1979; may have 1979 races not reported in 2002 and 2002 races not reported in 1979.

Table 7: Percent of Respondents for Whom Race Indicators Are in Agreement, by 2002 Self-Reported Race (non-Hispanics only)

2002 self-reported race (RACE02)	Report 1 race in 1979 and 2002		Report >1 race in 1979 and/or 2002	
	Num.	Percent Agree[a]	Num.	Percent Agree[b]
1-race categories				
White	2699	76.7	1102	93.7
Black	2173	96.7	88	89.8
Asian/Pacific Islander	21	66.7	7	42.9
American Indian	21	76.2	18	94.4
Other	37	5.4	7	0.0
Refuse	2	0.0	0	
All 1-race categories	4953	84.9	1222	92.6
All 2-race categories			70	28.6
All 3-, 4-race categories			8	0.0
Total	4953	84.9	1300	88.5

[a]The same race is reported in 1979 and 2002, and RACE78=black if the 1979/2002 race is black or else RACE78=other.

[b]All 2002 races are reported in 1979, and RACE78=black if black is reported in 2002 or else RACE78=other.

Table 8: Percent of Variance in Log-Wages Explained by Race and Ethnicity, Using Alternative Classification Schemes

Spec.	Description of race/ethnicity classification	R^2	Pct. change in R^2 from row 1
1	H,B,all other—using RACE78	3.8161	—
	Control for ethnicity only		
2a	H,non-H—using RACE78	0.0289	-99.2
2b	using RACE79	0.0298	-99.2
2c	using HISP02	0.0300	-99.2
	Control for race only		
3a	B,non-B—using RACE78	3.1760	-16.8
3b	using RACE79	3.0581	-19.9
3c	using RACE02	3.3983	-11.0
4b	W,non-W—using RACE79	2.5345	-33.6
4c	using RACE02	3.5056	-8.1
	Control for race only, using RACE02		
5	W,B,all other	3.8464	0.8
6a	W,B,A,AI,O,refuse, 1 aggregate multiple-race category	3.9719	4.1
6b	W,B,A,AI,OR, 1 aggregate multiple-race category	3.9673	4.0
7a	W,B,A,AI,O,refuse (minority assignment)	4.0274	5.5
7b	W,B,A,AI,OR (minority assignment)	4.0103	5.1
8a	W,B,A,AI,O,refuse, 12 separate multiple-race categories	4.0661	6.6
8b	W,B,A,AI,OR, 12 separate multiple-race categories	4.0616	6.4
	Interact race and ethnicity, using RACE02 and HISP02		
5'	(W,B, all other)xH	4.0017	4.9
6b'	(W,B,A,AI,OR, 1 aggregate multiple-race category)xH	4.1669	9.2
7b'	(W,B,A,AI,OR)x H (minority assignment)	4.2132	10.4
7b"	(W,OR)xH, B,A,AI (minority assignment)	4.1765	9.4
8b'	(W,B,A,AI,OR, 12 separate multi-race categories) xH	4.2751	12.0
8b"	(W,OR)xH,B,A,AI, 12 separate multi-race categories	4.2207	10.6
	Control for inconsistent responses using RACE02, RACE79, HISP02		
7b-i	W-C,W-I,B,A,AI,OR (minority assignment)	4.2412	11.1
7b"-i	(W-C,W-I,OR)xH, B,A,AI (minority assignment)	4.2837	12.3

H=Hispanic
W=white
B=black
A=Asian/Pacific Islander
AI=American Indian
O=other
OR=other and refuse combined
(…)xH shows variables that are interacted with Hispanic
W-C=white; 1979 and 2002 reports are consistent
W-I =white; 1979 and 2002 reports are inconsistent

Table 9: Log-Wage Variance Within Race/Ethnicity Categories for Selected Classification Schemes Shown in Table 8

Race/ethnicity category	Sample share	Within-category variance as percent of total variance, weighted by sample share					
		5	5'	7b	7b"	7b-i	7b"-i
White	.584	59.8		59.8			
Non-Hispanic	.503		51.9		51.9		
Hispanic	.081		7.8		7.8		
White (consistent)	.434					45.3	
Non-Hispanic	.420						43.5
Hispanic	.014						1.8
White (inconsistent)	.149					14.3	
Non-Hispanic	.083						8.3
Hispanic	.066						5.9
Black	.292	25.2		25.4	25.4	25.4	25.4
Non-Hispanic	.289		24.9				
Hispanic	.003		.4				
Asian	.006			.6	.6	.6	.6
American Indian	.014			1.4	1.4	1.4	1.4
Other	.125	11.1					
Non-Hispanic	.025		2.3				
Hispanic	.100		8.8				
Other/refuse	.104			8.9		8.9	
Non-Hispanic	.007				.5		.5
Hispanic	.096				8.4		8.4
Sum over categories ($1-R^2$)		96.2	96.0	96.0	95.8	95.8	95.7

Note: The number heading each column corresponds to the specification shown in table 8.