NISS

Multiple Imputation of Race for Project TALENT

Alan F. Karr

Technical Report 187
April 2014

National Institute of Statistical Sciences
19 T.W. Alexander Drive
PO Box 14006
Research Triangle Park, NC 27709
www.niss.org

Multiple Imputation of Race for Project TALENT

Alan F. Karr
April 21, 2014

1 Summary

This document describes a procedure for multiple imputation of race for all 377,015 students in the primary PT student datafile. The procedure incorporates not only information in this file, but also family information derived by AIR. It employs school-level racial composition information derived from this file, and also uses school-reported information in the General School Characteristics (GSC) datafile. Specifically, it employs the responses to questions GSC95–GSC99.

As discussed in §3, the procedure correctly predicts race for more than 95% of the 1952 respondents in the 2012 Pilot Study.

2 Details

Step 0: Racial Categories Race is imputed as *one of* "Asian," "Black," "White" and "AllOther." The rationale for this is that reporting of race is not consistent across the 1-year, 5-year and 11-year PT followups. Moreover, reported numbers in other categories (e.g., American Indian/Alaska Native and Hawaiian/Other Pacific Islander) are too small to support reliable imputation.

Step 1: Followup Race The procedure begins with creation of a "Followup Race" variable for all students, as follows:

1. If the student reports race in any of the followups, "Followup Race" is the most recently reported value.
2. Otherwise, Followup Race is set to missing.

Figure 1 shows the distribution of "Followup Race," which is missing for 202,504 of the 377,015 students (53.7%).

Step 2: Family Race The AIR-reconstructed "FamilyID" variable is present for 88,123 students, comprising 42,426 families. For these students, the "Followup Race" variable was (possibly) revised using the following logic: if all siblings in the family who reported a race reported the same race, that race was assigned to any sibling in the family who did not report a race.[1] This step constructs a race for 25,151 of the 42,426 families comprising 52,684 students. The remaining 35,439 students with a "FamilyID" cannot be assigned a race via this step.

[1] This reasoning is inconsistent with "modern" interpretations of race, but less inconsistent with interpretation of race in 1960.

Step 3: Racial Composition of Schools For each of the 1226 schools in the PT sample, the following were calculated from the student datafile:

1. The percentage of students with a nonmissing value of "Followup Race;"
2. The percentages of students, among those with a nonmissing value of "Followup Race," who reported their race as "Asian," "Black," "White" and "AllOther."

Step 4: Imputation of Race for Students with a FamilyID This imputation was carried out at the family level, i.e., for each of the 17,251 = 42,426 - 25,151 families for which a race was not assigned in Step 2, and the imputed family race was assigned to each student in that family, in the following manner:

1. If the percentage of students in a student's school who reported a race *exceeds 50%* (see §4 for discussion), then the family race is imputed randomly using the reported percentages calculated in Step 3.
2. Otherwise the race is imputed random using an alternative set of probabilities derived from the school's responses to questions GSC95–GSC99 in the following manner.

   These questions request percentages of students who are "Spanish or Latin American," "Oriental," "American Indian," "Negro [*sic*]," and "Other 'Minority' Group (specify)." The responses are categorial, consisting of "None," "00–9%," . . ., "90–99%" and "All." The "Other 'Minority' Group" category was ignored, because it is clear that some schools interpreted this to mean religious minorities in addition to or instead of racial minorities. To create consistency, the "Spanish or Latin American" and"American Indian" were merged into an "AllOther" category.

   The categorical responses were converted to numerical values using the range midpoints: 0, 5, . . ., 95, and 100, and then converted to probabilities: 0, 0.05, . . ., 0.95, and 1.0. When these probabilities summed to more than 1.0, which is possible because they are effectively categorical, they were rescaled to sum to 1.0. Finally, the probability of "White" is 1.0 minus the sum of the other probabilities.

Step 5: Imputation of Race for Students without a Family ID This imputation is performed exactly as in Step 4, but at the student level.

Five implicates were created. Figure 2 shows the distribution of imputed race for each. There is no significant implicate-to-implicate variability. A clear and important difference between the distribution in Figure 1 and those in Figure 2 is that in the former the percentage of blacks is only 4.5%, in each implicate this percentage is approximately 7.6%. In 1960, approximately 10.6% of the US population was black. Although there is some evidence that blacks are under-represented in PT, the implicates seem more plausible than the "Followup Race" variable alone.

3 Validation

An informal validation of the imputation was performed by comparing imputed races to races reported in the 1,952-respondent 2012 Pilot Study.

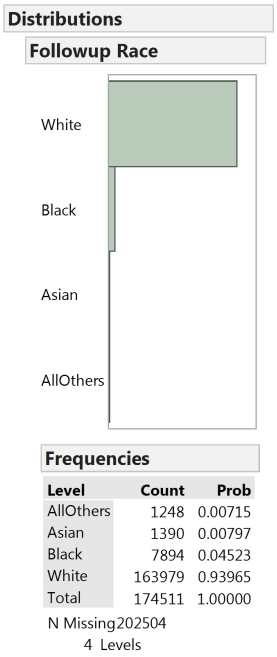

Figure 1: Distribution of the "Followup Race" variable prior to either family-derived assignment or imputation.

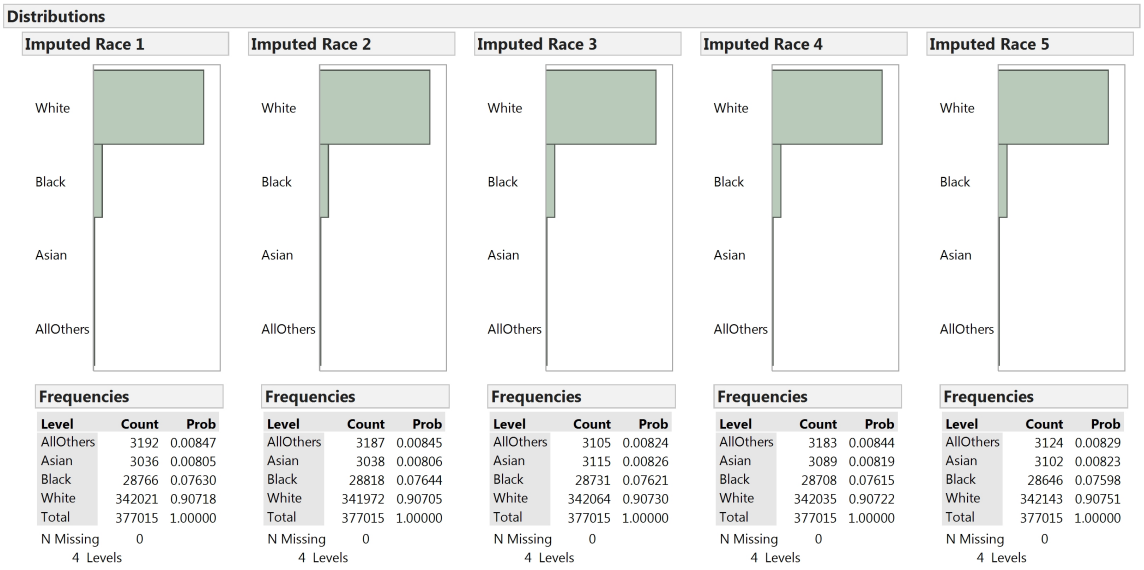

Figure 2: Distribution of imputed race in five implicates.

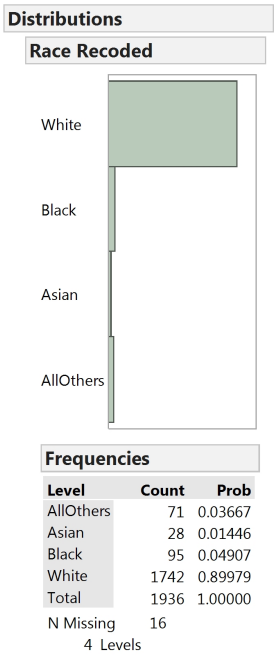

Figure 3: Distribution of recoded race in the 2012 Pilot Study.

From that study, a recoded, four-category race variable was created that is, to some extent, comparable to the variable used in the imputation; its distribution is shown in Figure 3.[2]

Race was imputed for 725 of the participants in the pilot study. Figure 4 shows that for all five implicates with imputed race is correct in approximately 96% of these cases. "Correct" means that the imputed race is among the races reported in the Pilot Study.[3] (The cases in Figure 4 labeled by "?????" are those for which no race is reported in the Pilot Study, meaning that correctness cannot be determined.)

A more nuanced, but consistent view is shown in Figure 5. In it, the cases are split by into those for race was imputed, which are labeled by "Race Imputation Flag = 1," and those for which race was not imputed, labeled by "Race Imputation Flag = 0." That the accuracy is not 100% in the latter arises from inconsistencies between the "Followup Race" and the race(s) reported in the Pilot Study. Accuracy for the races that are imputed ("Race Imputation Flag = 1") is on the order of 92% for all implicates, which is quite good.

4 Discussion

There is one settable parameter in the procedure described in §2: the threshold, which is 50% there, that determines whether student-reported or school-reported race composition is used to define the probabilities

[2]The lack of comparability results from recoding multiple race responses, of which there were 44—-approximately 5% of the data, to "AllOther," Since many multiple-race responses were black and another race, this recoded variable understates the percentage of blacks in the data.

[3]There is no meaningful possibility of imputing multiple races from followup data, because there are no relevant data.

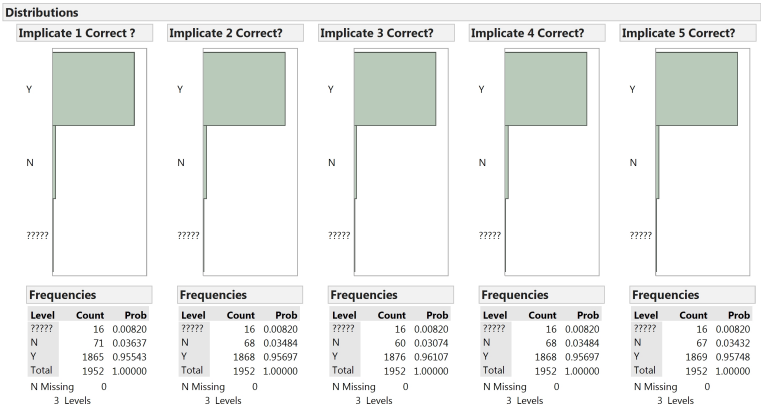

Figure 4: Distribution of recoded race in the 2012 Pilot Study.

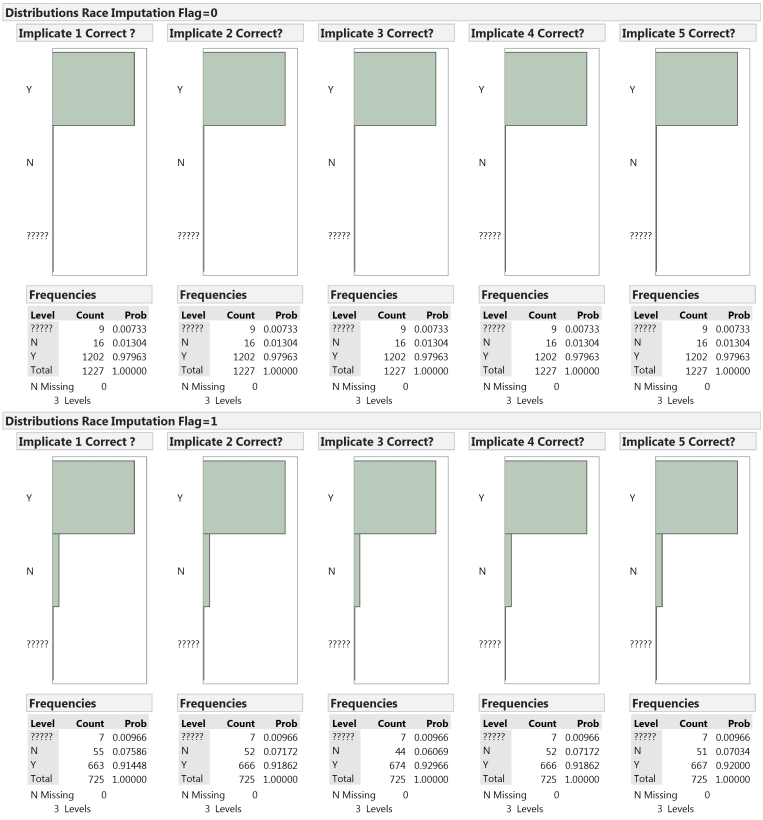

Figure 5: Distribution of recoded race in the 2012 Pilot Study.

employed in the imputation. An informal sensitivity analysis shows that the value of this parameter does not have major effects.

It is important to bear in mind that the purpose of the imputation is to provide race for those students in the PT sample, not to reproduce a national- or state-level racial composition. The percentage of participants in the 2012 Pilot Study who reported their race as "Black" (and possibly something else) is 5.1%, which is below the percentage of the population that was Black (from §2, 10.6%). The imputed percentages of Blacks are closer to this latter value, but may be too high.[4]

What does seem clear is that there is a major under-representation of Blacks in PT, especially in some states. For instance, the percentage of students in Mississippi whose "Followup Race" is black is only 2.2%, which is not remotely close to the percentage of the 1960 population of Mississippi that was Black, which is 42%. This raises the obvious issue of whether there should be a nonresponse adjustment to the base year PT weights for race bias.

[4]It is almost impossible to impute anyone's race as Black without "over-imputing."