Using Names To Check Accuracy of Race and Gender Coding in NAEP

Jennifer Czuprynski Kali, James Bethel, John Burke, David Morganstein, and Sharon Hirabayashi
Westat

Keywords: Data quality, Coding errors, Mapping

1. Introduction

The National Assessment of Educational Progress (NAEP), also known as 'The Nation's Report Card', is an assessment of student performance conducted bi-annually by the National Center for Education Statistics (NCES). The assessments are conducted on a sample of fourth, eighth, and twelfth grade students on various subjects, including reading, mathematics, and science.

The sample is a two-stage design, in which students are sampled within sampled schools. In 2005, over 1.2 million students were sampled from over 20,000 schools. Because of the size of the study, an electronic filing (e-filing) system was developed by Westat to obtain the sampling frames for the second-stage of sampling in an effort to automate the sampling process. Sampled schools are asked to upload spreadsheets containing all students in the sampled grade, with corresponding demographic variables that are used in the second-stage sampling process, such as race/ethnicity and gender.

The variables in the spreadsheets come in various orders and with many variations on variable names (such as race, ethnicity, ethnic code, race/ethnicity, etc.) The values in each of the variable columns vary with each school as well; for example a school might use letters, numbers, or the words Black, White, etc., to indicate race/ethnicity. To understand the data on the spreadsheets, schools are asked to map the columns and values in the columns to specific values and variable names.

On several occasions, the agencies reporting these data have incorrectly mapped the race and/or ethnicity of the students being tested. Historically, when these data were reported by hand, such errors were on a small scale. However, NAEP has grown extensively and the advent of e-filing greatly expands the potential effects of such errors. The race/ethnicity coding of an entire school district or even state may be incorrect.

Such errors can have serious consequences and yet are difficult to detect. This paper discusses statistical checks to uncover these errors. The first check, which is administered online in real time, compares percentages in each demographic group against known percentages from government databases, such as the Common Core of Data (CCD). However, a percentage check will not detect the error of a school incorrectly mapping two groups that are similar in size (e.g., male and female students, or small ethnic groups).

Given the importance of classifying students in the correct demographic groups, and the inability to detect mapping errors based on percentages alone, we developed other methods for checking the data classifications. This paper describes the approach we developed, which uses data on frequencies of names for given race/ethnicity and gender classifications to determine possible mapping errors. The method examines the question, 'did the person who submitted a school list correctly map all students' gender and race?' The method does not examine individual students' characteristics; rather, it is based on joint probabilities computed for all students in a submission. What is more, it uses first names and last names separately and complete names of students are not included in the databases described below.

2. Developing Checks Based on Names

2.1 Creating a Historical Database of Names

Historical databases of first names and last names were created from the schools that e-filed during the NAEP 2003 and 2004 assessments. These databases were developed from over one million raw records, 1,053,772 from 2003 and 103,391 from 2004. This list was reduced to 1,130,764 records after removing records with improper format, such as quotation marks or other symbols in the name field, and records from Hawaii.[*]

There were 162,549 unique last names and 89,476 unique first names. For each first name, we computed (1) the probability that a student with the name is a given race/ethnicity for each possible race/ethnicity (i.e., six separate probabilities), (2) the probability that a student with the name is a given gender for both male and female (two probabilities), and (3) the probability that the name is a first name. For each last name, we computed the probability that a student with the name is a given race/ethnicity and the probability that the name is a last name. In all cases, the probability calculations used this formula:

[*] From prior experience, it was known that data on race/ethnicity and ethnicity from Hawaii were complex and not representative of the rest of the United States. In addition, the relationship between name and race/ethnicity in Hawaii seems to be unique in the United States and it was felt that these data would not be helpful in formulating a method that would be applied to the United States at large.

$$p_i(r) = \frac{\text{Number of students on database with name } i \text{ and } r}{\text{Number of students on database with } r}$$

where r is the corresponding classification, either for race/ethnicity, gender, or the entire file for the first and last name checks.

Note that the historical database is not a perfect set of names, race/ethnicities, and genders. We cleaned the database as much as possible, utilizing knowledge of race/ethnicity and gender data problems in 2003 and 2004. Although it is not possible to create an errorless file, we believe the effect of the remaining errors is minor.

2.2 Reverse Likelihood Function

The scoring procedure used to create the probability of a list of students belonging to a given category uses a variation of the likelihood function. In most applications of the likelihood function, probabilities for a series of independent observations are multiplied together. The function is then maximized over a set of parameters. There is a key difficulty in applying this approach to the names problem. Most names occur infrequently, so that many names would have $p_i(r) = 0$. As a result, the likelihood function is not useful for this problem. As described below, we circumvented this problem by reversing the usual formulation of the likelihood function.

The "reverse" likelihood function (RLF) is calculated separately for first and last names and for each race/ethnicity and for each gender on the incoming file against one or more of the races on the historical file.

For a given category r (race/ethnicity, gender, or name order) on the incoming file, the RLF for historical r would be calculated as:

$$\begin{aligned}\text{RLF}(r) &= -\ln\left(\prod_{i=1}^{n}\left(1-p_i(r)\right)\right)\\ &= -\sum_{i=1}^{n}\ln\left(1-p_i(r)\right)\end{aligned}$$

where r is as defined above, i refers to each name on the list, and n is the total number of names in the specific category on the e-file.

In this formula, duplicate names on the e-file (e.g., multiple students with the name "Smith") are all counted in the calculation formula. Notice that by calculating $(1 - p_i(r))$, we are finding the probability that the name i does *not* belong to the given race/ethnicity, thus reversing the usual likelihood function. The purpose of this is to avoid the many instances where rare names cannot be matched to the historical file. By changing the sign (i.e., taking the negative of the logarithm), maximizing the RLF may be interpreted as minimizing the probability that a given name is *not* typical of the race/ethnic group being considered. In other words, maximizing the RLF serves the same purpose as maximizing the usual likelihood function.

Tables 1-3 show illustrative RLF scores that were computed using all of the student names e-filed in NAEP 2005 as compared with the historical file for race/ethnicity. Table 1 shows the RLFs for last names, Table 2 shows the RLFs for first names, and Table 3 shows the RLFs for first and last names combined (computed by summing the RLFs for first and last names).

In Table 1, notice that the last names of Black, non-Hispanic students in the incoming file have an RLF score of 150 when compared to students with White, non-Hispanic names on the historical file versus a score of 421 for the Black, non-Hispanic last names on the historical file. The last names for Black, non-Hispanic names show the highest scores when compared to Black, non-Hispanic names on the historical file; similarly, Hispanic names score highest when compared with Hispanic names and Asian/Pacific Islander names score highest when compared with Asian/Pacific Islander names. Only White, non-Hispanic and Native American/Alaskan Native names do not.

Table 2 indicates that first names of White, non-Hispanic students score highest when compared with first names of White, non-Hispanic students in the historical file. However, Native American/Alaskan Native students do not appear to have distinctive first or last names, since their RLF scores are not highest in either Table 1 or 2 when matched against Native American/Alaskan Natives in the historical file.

Table 3 shows the RLF scores when first and last name are combined. Since the RLF consists of summing logarithms, the combined scores are the sum of the scores for the first and last names. The combined RLFs have the advantage that White non-Hispanic, Black non-Hispanic, Hispanic, and Asian/Pacific Islander names all score highest when compared against their own names in the historical file. This combined RLF was used in the screening algorithm to check for mapping errors.

The associations are much stronger for gender, using only the first names.

2.3 Developing the Screening Algorithm

The screening algorithm utilizes the RLF to predict the probable race/ethnicity, gender, or name order of a given list of students. This section describes the steps in using RLF scores to develop the screening algorithm.

2.4 Race/Gender Check Algorithm

Schools provide a file containing student name, gender, and race/ethnicity data. Student name is separated in two fields—first and last. Gender and race/ethnicity are uploaded as one field each. The two student name fields are required, but neither gender nor race/ethnicity is required. There are two possible values for gender (male, female) and six possible values for race/ethnicity (White, non-Hispanic; Black, non-Hispanic; Hispanic; Asian/Pacific Islander; Native American/Alaskan Native; Other). A file may have 0-2 values for gender, and 0-6 values for race/ethnicity.

We want to determine if the person submitting the file, later referred to as the e-filer, correctly designated the entire collection of students on the file as male or female, and into the correct racial and ethnic categories. Since our check is dependent upon the students' first and separately last names, it is also important that the e-filer designated the correct columns for first names and last names. Since many last names are male first names (Stanley, Douglas, Mitchell, etc.), if the last name is designated as the first name, both groups of students appear to be male students.

Instead of determining if the individual designations are correct, we focus on determining the most likely classifications. Therefore, the algorithm is not concerned with which group of students on the file are labeled by the e-filer as 'Male'. The algorithm calculates a joint probability to determine which group of students is most likely the 'Male' group. Therefore, the groupings of the e-filer are preserved, but not the labels for the groups. In other words, the students labeled as male are assumed to be of the same gender, and the algorithm predicts the gender for that group using a joint probability for the set. The assumption is that the e-filer correctly grouped their students according to gender and race/ethnicity, but may have mislabeled the group.

The likelihood of each possible permutation is computed by comparing the names on the uploaded file to the names in the historical database. A file is required to have both first and last names, but may have 0-2 values for gender, and 0-6 values for race/ethnicity. Therefore, the number of possible permutations is:

$$2 \cdot g!\binom{2}{g} \cdot r!\binom{6}{r}$$

where g is the number of gender values on the file and r is the number of race/ethnicity values on the file. If a file contains both genders and all six race/ethnicity categories, there are 2,880 possible permutations.

Note that for race/ethnicity and gender, the RLF is computed for each category within each first or last name file, but for the name order check, the RLF is computed for each category on the entire first and last name files.

2.5 Reverse Likelihood Function for All Permutations

The likelihood of a given ordering is computed simultaneously for up to ten possible outcomes: name order (first or last), gender (male or female), and race/ethnicity (White, non-Hispanic, Black, non-Hispanic, Hispanic, Asian/Pacific Islander, Native American/Alaskan Native, or Other). The algorithm cannot detect a situation where more than one grouping corresponds to the same classification (e.g., both Asian/Pacific Islander and Black, non-Hispanic students mapped to Hispanic). We assume that the e-filer has grouped the students correctly, but may have mislabeled them.

To see how the RLF is computed for a permutation of characteristics, consider the following example looking at only the race/ethnicity mappings. Suppose that a school reports scores for four race/ethnic groups: r_1, r_2, r_3, and r_4. The reverse likelihood function can be computed for all possible permutations of choosing four races out of six (360 possible permutations). The general formula for each permutation is as follows:

$$\begin{aligned}\mathrm{RLF}(r_1, r_2, r_3, r_4) &= -\ln\left(\prod_{k=1}^{4}\prod_{i=1}^{n}\left(1-p_i\left(r_k\right)\right)\right) \\ &= -\sum_{k=1}^{4}\sum_{i=1}^{n}\ln\left(1-p_i\left(r_k\right)\right)\end{aligned}.$$

To compute the likelihood that the four submitted race/ethnicity groups, in order, are White, non-Hispanic (W), Black, non-Hispanic (B), Hispanic (H), and Asian/Pacific Islander (A), using the data in calculations in Table 3:

RLF($WBHA$) = 1,502.43 + 568.08 + 974.89 + 128.54 = 3,173.95

The RLF for White, non-Hispanic (W)-Hispanic (H)-Black, non-Hispanic (B)-Asian/Pacific Islander (A) would be

RLF($WHBA$) = 1,502.43 + 125.23 + 114.09 + 128.54 = 1,870.30

Notice that RLF(*WHBA*) is substantially smaller than RLF(*WBHA*), indicating that between these two choices, the original coding appears to be the correct one. This process would be repeated for all 360 possible permutations to find the most likely.

The same idea can be applied to gender and name order. Combining the three checks—name order, gender, and race/ethnicity—into one permutation allows us to check race/ethnicity and gender regardless of whether the name order is correct. Computing the joint permutations is simply a matter of adding the RLF's for each of the combinations of race/ethnicity and gender within each ordering of the names.

The general formula of the reverse likelihood (RLF) for a given permutation of *k* characteristics, each one taking on p_i alternative values, is given by the formula

$$\mathrm{RLF}\left(c_{11}, c_{12}, \ldots, c_{1p_1}, c_{21}, c_{22}, \ldots c_{2p_2}, \ldots c_{k1}, c_{k2}, \ldots c_{kp_k}\right)$$

$$= -\ln\left(\prod_{r=1}^{k}\prod_{l=1}^{p_r}\prod_{i=1}^{n}\left(1 - p_i\left(c_{rl}\right)\right)\right)$$

$$= -\sum_{r=1}^{k}\sum_{l=1}^{p_r}\sum_{i=1}^{n}\ln\left(1 - p_i\left(c_{rl}\right)\right)$$

2.6 False Positives

In developing the algorithm, reducing the false positive rate was one of the main concerns. Because we developed and tested this algorithm on the NAEP 2003 and 2004 datasets, which themselves contained unknown errors, it was not possible for us to separate true false positives from true errors in the datasets. We took steps to correct known errors on the NAEP 2003 and 2004 files, but beyond that, it is not possible for us to determine if an error found by the algorithm is a true error or a false positive.

We considered deleting names from the historical database that only occurred once, which would reduce the impact of header rows that were accidentally included on the file. Certain names on the file were FEMALE and CODE. While the group of names that occurred only once was relatively small overall (5.8% of first names and 8.9% of last names), Asian/Pacific Islander names were overrepresented within this group (16.7% of Asian/Pacific Islander first names and 17.2% of Asian/Pacific Islander last names occurred only once). Therefore, not including these names would reduce our ability to detect groups of Asian/Pacific Islander students. Thus, while this may be a good idea for the future, when there are more names in the historical database, it was not feasible in 2005.

Instead of focusing on false positives, we took steps to reduce the total number of errors. We controlled the overall error rate by setting a threshold for the number of students required to perform the check. It was decided that each school, and each category of students within a school, would need at least ten students in order to be checked. Therefore, the check is not performed for any schools with less than ten students. If a school has more than ten students, but an individual category does not, then that category alone will not be included in the permutation. For example, if a school has 15 students: 10 male, 5 female, 11 White, non-Hispanic, 4 Black, non-Hispanic; the permutations will be based only on the 10 male and 11 White, non-Hispanic students. Possible permutations would be (first, last, male, White, non-Hispanic), (first, last, female, White, non-Hispanic), etc.

3. Implementation and Results

In NAEP 2005, electronic lists could either be submitted for individual schools or simultaneously for multiple schools in a single file, such as an entire state or school district. The files are run through a series of checks while the user is online to verify that the data was correctly uploaded. The race/ethnicity and gender checking algorithm was run on submitted files after the data had been verified by the user following the online checking process, meaning that the data checked in the race/gender algorithm has had some initial quality checking.

Multiple groupings were checked at both the file level and the individual school level, hoping to identify problems related to mapping errors in the uploading process, and problems related to states or districts putting together multiple school files with varying coding schemes.

A team of statisticians and programmers reviewed all failures to separate out the false positives from the actual failures. A red light system was created: a coding of green indicated a false positive, a coding of red indicated a definite failure, and a coding of yellow was applied to cases that were neither green nor red. Files with a green light were sent through the sampling process. Files with a yellow light were also sent through the sampling process, but a note was sent to the field staff conducting the assessments to look for errors. For multiple jurisdictions with yellow lights at the file level, the field staff was notified about all schools in the file. For red lights, the e-filer was contacted to resolve the error before samples could be drawn for that school. Any red light schools remaining at the end of the e-filing period were treated as yellow lights, sending the file through the sampling process with a note to the field staff regarding the possible error.

The number of failures and their green, yellow, or red status are shown in Table 4. Contact was made with the e-filer for 33 of the 57 red lighted single schools, and new files were uploaded with the errors corrected. The remaining 24 were not able to be contacted, and thus were treated as yellows and a note was sent to the field staff assessing the school. The one red multiple file was resolved by contacting the e-filer. For the single schools within multiples, two of the red lighted schools were resolved by contacting the e-filer, and the remaining five were treated as yellows.

The final status of the yellow schools is yet to be determined. We need to analyze data from the field to determine if these resulted in actual errors that the field staff corrected, or if the field staff found that the data was correct.

The human review found a large number of false positives. Though the overall failure rate was relatively low (6.75%), the majority of these were found to be false positives. The gender check produced few failures overall, and very few of those were false positives, so the discussion of false positives focuses on the race/ethnicity check.

We believe the large number of false positives is mainly due to an inadequate database of names for some of the smaller ethnic groups: Pacific Islander students in Hawaii, Native students in Alaska, and Native American students in the West. This problem may correct itself with the larger list of names that will be added from the 2005 database. However, it may be beneficial to supplement our database of NAEP students with common names from each of the ethnic groups listed above.

Another issue deals with the diversity of names within each ethnic group. Names of students in certain smaller ethnic groups within larger race/ethnicity classifications seem to correspond with different classifications. For example, Filipino students, mainly in California, are classified as Asian/Pacific Islander, but have names which correspond to more to Hispanic names than to Asian/Pacific Islander names. We do not think there is a way to correct for this, and thus regardless of the amount of tweaking, there will always be some degree of false positives in this algorithm.

A possible cause of the problem with Native American names may be the way the names are treated in the algorithm. Names with more than one word are dropped to only the first word for the historical database and also for the school lists. Therefore, we are not capturing the Native American naming conventions, which often have last names like Yellow Bird and Standing Tall, possibly the reason that the algorithm is not able to correctly identify Native Americans.

Another large group of race/ethnicity failures corresponds to small schools, mainly in the West, which are 100% White, non-Hispanic. Often, the check would find these schools to be more likely 100% Black, non-Hispanic. There is also a large number of 100% Native American/Alaskan Native schools which, according to the check, were more likely White, non-Hispanic or Black, non-Hispanic. It would be useful to suppress these errors, possibly by comparing the racial percentages to CCD percentages before setting the error flag.

4. Discussion

4.1 Summary

We developed a race/ethnicity and gender check based on students first and last names in the submitted files. For each school, the check compares the first and last names of a group of students within race/ethnicity and gender categories to a historical database of first and last names. A unique first name has a probability that it is a male or female name based on the frequencies of the name on the historical file. Similarly, each first name and last name separately has a probability attached to each race/ethnicity category. Based on the probabilities in the historical database, the most probable race/ethnicity and gender categories for the file submitted for the school are determined. The school fails the check if the most probable race/ethnicity and gender categories do not match the categories submitted by the school.

Because this check looks at the entire school as a whole, and not at individual students, it is designed to identify labeling errors—meaning the school has each student assigned to the correct race/ethnic or gender group, but the mapping from the school codes to the NAEP codes was done incorrectly. This check is not designed to determine if the students are correctly grouped.

Overall this algorithm was a success, enabling us to find and correct 36 errors in files in-house, and give field staff warnings to 106 others.

4.2 Suggestions for Changes

As mentioned previously, our database is lacking in some of the smaller ethnic groups, reducing our ability to recognize these ethnic groups. It may be useful to supplement the database of names, though it may be sufficient to add the student lists from NAEP 2005 to the over one million names already in the historical database.

It is necessary to address the treatment of Native American names, since the naming structure of Native American names differs from most other ethnic groups. As it is, we are not able to recognize Native American names with this algorithm.

It may be beneficial to compare the suggested permutations to the racial percentages found in the CCD, as a way of fine tuning the check and reducing the number of false positives.

Looking into a test of significance, such as a likelihood ratio test will be beneficial in reducing the number of false positives.

This is a very time-intensive check, with human review of every failed school or jurisdiction, for over one thousand failures.

Implementing some of the changes listed here may reduce the errors, but it would also be beneficial to determine ways to automate more of the process.

Starting in NAEP 2009, schools will be able to report data on students with multiple race/ethnicity categories. Our current algorithm is not able to handle this change and must be updated in order for the race/ethnicity checks to remain relevant.

Table 1. RLFs for last names

	Race/ethnicity on historical file					
Race/ethnicity on incoming file	White, non-Hispanic	Black, non-Hispanic	Hispanic	Asian/ Pacific Islander	Native American/ Alaskan Native	Other
White, non-Hispanic	**191**	**396**	42	42	188	196
Black, non-Hispanic	150	**421**	19	34	162	183
Hispanic	13	17	**626**	26	49	78
Asian/Pacific Islander	4	8	10	**93**	7	10
Native American/ Alaskan Native	4	**10**	4	1	**6**	5
Other	2	**6**	4	2	3	3

Table 2. RLFs for first names

	Race/ethnicity on historical file					
Race/ethnicity on incoming file	White, non-Hispanic	Black, non-Hispanic	Hispanic	Asian/ Pacific Islander	Native American/ Alaskan Native	Other
White, non-Hispanic	**1,311**	519	468	503	900	880
Black, non-Hispanic	**193**	**146**	94	89	156	163
Hispanic	200	107	**348**	119	155	171
Asian/Pacific Islander	**60**	28	30	**34**	44	44
Native American/ Alaskan Native	**21**	9	8	8	**16**	15
Other	**13**	6	6	6	10	10

Table 3. RLFs for combined first and last names

	Race/ethnicity on historical file					
Race/ethnicity on incoming file	White, non-Hispanic	Black, non-Hispanic	Hispanic	Asian/ Pacific Islander	Native American/ Alaskan Native	Other
White, non-Hispanic	**1,502**	916	511	546	1,089	1,077
Black, non-Hispanic	343	**568**	114	124	318	347
Hispanic	214	125	**975**	146	204	249
Asian/Pacific Islander	65	37	41	**129**	51	54
Native American/ Alaskan Native	**28**	20	13	11	**22**	21
Other	**16**	13	11	9	13	14

Table 4. Summary of race/gender failures and status codes

		Multiple school listings	
	Single schools	Entire file	Single schools w/in file
Total files	**6,325**	**79**	**9,310**
Total failures	**368**	**25**	**652**
Gender failures	17	0	2
Green	1	--	1
Yellow	0	--	1
Red	***16***	***--***	***0***
Race/ethnicity failures	348	25	649
Green	275	20	380
Yellow	34	2	40
Red	***39***	***1***	***7***
Both gender and race/ethnicity	3	0	1
Green	1	--	1
Yellow	0	--	0
Red	***2***	***--***	***0***