NAGB Conference on Increasing the Participation of SD and LEP Students in NAEP

Commissioned Paper Synopsis

The attached paper is one of a set of research-oriented papers commissioned by NAGB to serve as background information for the conference attendees. The authors bear sole responsibility for the factual accuracy of the information and for any opinions or conclusions expressed in the paper.

Including Students with Disabilities and English Language Learners in NAEP: Effects of Differential Inclusion Rates on Accuracy and Interpretability of Findings

Edward H. Haertel, Ph.D.

Stanford University

December 2003

- The paper initially describes the sources of uncertainty in NAEP data and standard errors. As NAEP sample sizes have increased (beginning in 2002), greater precision (smaller standard errors) has been attained by the program. For this reason, exclusion effects are increasingly important.

- Two scenarios of revised NAEP results are presented (for New York City and for the nation) that reflect the possible results if all excluded students had been included in the data analysis (one scenario presents results assuming that all excluded SD were assigned the mean score for all SD tested and another scenario presents results where SD were assigned a score two standard deviations below the mean for SD). The overall NAEP results from the two recalculation scenarios vary considerably. Even where exclusion rate is constant, exclusion may affect score comparisons. When exclusion rates are not constant over time, the effects of exclusions on data comparisons can be dramatic.

- NAEP results can be affected by the percentage of students identified as SD or LEP in states or districts, as well as exclusion rates. The paper presents estimated recalculated results for the Trial Urban District Assessments in the two scenarios above to show how the rank orders of the districts' performance might have changed substantially. Student subgroup results may also change dramatically with increased inclusion.

- The effects of exclusions on NAEP data reliability can be minimized (1) by minimizing exclusions; (2) by establishing exclusion criteria that are as clear and objective as possible and working to assure that those criteria are adhered to; and (3) making practices and criteria across states as uniform as possible

- Remedies for the general effect of exclusions: First, efforts to minimize exclusions should continue. The higher the proportion of exclusions, the greater the bias and the greater the uncertainty about the true performance level of the population. Second, NAEP users should perhaps be reminded more often that the students tested do not represent the entire population. Finally, research should continue on the utility of imputation models that might be used to adjust for effects of exclusions.

- The consequences of differential exclusion policies are serious. If such policies vary for the two time points, groups, or jurisdictions compared, then that first, fundamental inference is compromised. The observed change, contrast, or performance gap in fact represents some mixture of differences in actual student achievement distributions and differences in decision rules determining whom to test. The comparison of performance for tested groups is thus "apples to oranges." The increasing accuracy of NAEP statistics has made even small distortions more important than they once were.

- Overall, for 2002 fourth-grade reading, the correlation across states between the percent of all students *identified* as SD and/or LEP and the percent of all students *excluded* as SD and/or LEP was only .38. If policies were uniform, a higher correlation would be expected. For the 36 states that participated in both the 1998 and the 2002 fourth-grade reading assessment, the correlation between the *change* in exclusion rate from 1998 to 2002 and the *change* in achievement from 1998 to 2002 was .57. This indicates that a substantial proportion of variation in states' achievement score changes can be accounted for by changes in their rates of exclusion.

- For the most part, effects of exclusions on reliability can be offset by increasing sample sizes. Effects of exclusions on validity are more problematical. It is important that NAEP continues to keep exclusions to a minimum, and that efforts be made to work toward more uniform policies and practices for determining which students should be excluded from NAEP and which should be tested.

Including Students with Disabilities and English Language Learners in NAEP: Effects of Differential Inclusion Rates on Accuracy and Interpretability of Findings

Paper prepared for the National Assessment Governing Board

Edward H. Haertel, Ph.D.
December 2003

The National Assessment of Educational Progress (NAEP) provides the best information available for comparing student achievement across states and (recently) large urban districts, and for monitoring national performance trends over time. The unique value of NAEP comes from its standardization of tests and administration procedures for all participating states and districts. A typical state assessment may provide better information than NAEP about patterns of achievement in that one state in a given year, because it is likely to be more closely aligned to the state's own academic content standards and involves testing larger samples. In general, though, state assessments cannot support accurate comparisons of one state to another. For such comparisons, NAEP is unparalleled. NAEP also excels in maintaining uniform reporting scales over long periods of time. Most state assessments have changed tests, time of year of testing, grade levels tested, etc. much more frequently than NAEP. Any such changes are likely to disrupt trend lines.

The most obvious aspect of NAEP's standardization is the use of the same test booklets and administration procedures for all participating states and districts, but another critical aspect is the use of common definitions for the student populations sampled. If participating states or districts follow different practices for excluding students with disabilities (SD) or limited English proficient (LEP) students, that second kind of standardization is compromised.

The technical details are complicated, but in essence NAEP relies on two methodologies. First, NAEP relies on measurement theory (psychometrics). State-of-the-art methods are used to construct the framework for the assessment, write and field test items, and estimate score distributions on NAEP proficiency scales. Comparability requires uniform measurement procedures for all participating jurisdictions. Second, NAEP relies on statistical sampling theory. Samples of schools are chosen following complex procedures to assure their representativeness, and students at target grade levels are randomly chosen for testing within each participating school. Statistical theory is used to make inferences from the performance of the students tested to the populations from which they were sampled, and to quantify the accuracy of those inferences. Comparability also requires uniform population definitions and sampling procedures for all participating jurisdictions.

The numbers that NAEP provides, including both mean scores and percents at-or-above Basic or Proficient, etc., are widely interpreted as representing the performance of all students in the groups they refer to. Typically, these target groups (populations) include all students enrolled at a given grade level in a state or district's public schools, or subgroups of such students defined by gender, race/ethnicity, or other demographic variables.[1] In fact, however, NAEP has never characterized the performance of all such students. Exclusions of SD and/or LEP students change the definition of the populations assessed.

This paper reviews the implications of SD and LEP exclusions for the reliability and validity of NAEP trend lines, state and district comparisons, and measurements of gaps among subgroups. It begins with a brief discussion of statistical uncertainty in NAEP results, because the importance of any biases or distortions due to exclusions must be evaluated against the backdrop of the general uncertainty in NAEP results. Next, some effects of exclusions are illustrated with numerical examples. The discussion then turns to reliability and then to validity issues, followed by a brief conclusion.

Statistical Uncertainty in NAEP

Many sources of uncertainty affect the numbers NAEP provides, some random and some systematic. These include random errors affecting students' responses to test questions (carelessness and guessing), the random sampling of students within participating schools, the random sampling of schools within jurisdictions (sampling error), and many smaller sources of uncertainty.[2] When NAEP reports statistics such as mean scores or percentages of students at or above NAEP achievement levels, a *standard error* is also calculated for each statistic reported. The standard error summarizes the degree of uncertainty in the corresponding statistic.

For example, the average fourth-grade reading composite score in 2003 for all public-school students in New York City was estimated from the performance of 2,403 tested students. The estimated value on the NAEP score scale was 210. More precisely, the estimate was 209.88, with an estimated standard error of 1.39.[3] To understand what this number means, imagine that the NAEP assessment in New York City in 2003 could have been done not just once, but over and over again. Each time, different samples of students would have been chosen from different schools. The resulting estimates of the average fourth-grade reading score would not all have been exactly the same. Roughly speaking, a standard error of 1.39 means that the estimated mean score would be within 1.39 of the (unknowable) true, error-free population mean about 68 percent of the time, and within 1.96 x 1.39 = 2.72 of the true value about 95 percent of the time. As a practical matter, the standard error of 1.39 implies that in the number 209.88, no confidence should be placed in the hundredths digit or even the tenths digit. These extra

[1] At the national level, NAEP also includes private-school samples. Additional samples defined by age instead of grade level are used for long-term trend reporting. Other special samples (e.g., the charter school sample) are also drawn from time to time.

[2] Other sources of error include the random assignment of different test questions to different students under NAEP's "matrix sampling" design, refusal of some sampled schools or students to participate, student absences, imperfections in the lists of all schools from which the NAEP samples are chosen, imperfections in the lists of students at tested grade levels within participating schools, and, for constructed-response items, scorer error.

[3] All NAEP statistics in this paper were obtained or calculated using either the NAEP Data Tool at http://nces.ed.gov/nationsreportcard/ or data from published NAEP reports downloaded at the same website. Further details on exact sources and calculations are available upon request from the author.

decimal places are useful at intermediate stages in statistical calculations, but not for reporting. A 95 percent confidence interval (an interval that would cover the true value 95 percent of the time) can be constructed by taking 209.88 plus or minus 1.96 standard errors, in this case an interval from 207.16 to 212.60. Our best estimate of average fourth-grade reading proficiency in New York City is 210, but the standard error tells us that an estimate of 209 or 211 would be almost as good, and an estimate of anywhere from 207 to 212 or 213 is a reasonable possibility.

Most of the sources of uncertainty in NAEP statistics diminish as the number of students tested increases. For example, the average fourth-grade reading composite score in 2003 for all public-school students in the United States was estimated from the performance of 179,013 tested students, almost 75 times as many as were tested in New York City. The national average was estimated as 216.46, with a standard error of 0.27. A 95 percent confidence interval for the national mean, from 215.93 to 216.99, is considerably narrower than the confidence interval for New York City, 1.06 points versus 5.44 points. This shows the greater precision (smaller standard error) that comes with a larger sample size.[4]

Effects of Exclusions on NAEP Estimates

Table 1 illustrates the relevance of statistical uncertainty for understanding the implications of SD and LEP exclusions. (Average scores in Table 1 are shown to the nearest tenth of a point to show the effects of small changes.) The first three rows show the number of students tested, the average score, and the standard error for New York City and for the United States in 2002 and in 2003. The standard error in the "Change" columns is larger because the observed change is influenced by uncertainty in both of the scores compared. The one-year changes in performance were not statistically significant for either New York City or the Nation. The next three rows show the performance of students with disabilities who were tested as part of NAEP, with or without accommodations. Not surprisingly, performance of this subgroup is markedly lower in all cases than the performance of the group as a whole.

The next row (third from the bottom) shows what percent of all students sampled were excluded from the NAEP assessment due to a disability. The goal of NAEP is to include as many students as possible, but students with an IEP or those covered by Section 504 of the Rehabilitation Act of 1973 are excluded if the school's IEP team has determined the student could not participate, if the student's cognitive functioning is too severely impaired to enable participation, or if the student's IEP requires an accommodation or adaptation that NAEP does not allow and the student could not earn a score that accurately represented his or her abilities without that accommodation. As seen in Table 1, about one in twenty fourth grade students randomly chosen as part of the national NAEP samples in each of 2002 and 2003 were excluded under these criteria. In New

[4] Other things being equal, the standard error is roughly inversely proportional to the square root of the sample size. When comparing the Nation to New York City, this relation does not hold exactly because other properties of the respective samples differ.

York City, about one in twenty were excluded in 2002 and about one in fifty were excluded in 2003.

It is impossible to know how well the excluded students would have performed if they had been tested, but it is reasonable to assume that, on average, those excluded would have performed worse than those students with disabilities who were actually tested. The last two lines of Table 1 offer just two of the infinite set, of scenarios for the results that might have been obtained if these students had been tested. There is no entirely satisfactory way to "adjust" for the effects of exclusions. These illustrations merely present some possibilities for purposes of discussion. The first, conservative scenario assigns all students excluded for reason of disability the average score for tested students with disabilities. The bottom row assigns these students a score two standard deviations below the mean for tested students with disabilities.[5]

Table 1. Grade 4 Public School Reading Composite Scores, 2002 and 2003, for New York City and for the Nation.

	New York City			**United States**		
	2002	**2003**	**Change**	**2002**	**2003**	**Change**
Total Number Tested	862	2403		133805	179013	
Average score for all students	**206.2**	**209.9**	**3.7**	**216.8**	**216.5**	**-0.3**
Standard Error	2.6	1.4	2.9	0.5	0.3	0.5
Number of SD tested (approx.)	87	285		11040	17050	
Average score for students with disabilities	179.3	180.6	1.3	186.6	184.4	-2.2
Standard Error	5.41	4.04	6.8	0.81	0.60	1.0
Percent of all students sampled who were excluded due to a disability	5%	2%		5%	5%	
Average score if all excluded SD were assigned the mean score for SD tested	**204.8**	**209.3**	**4.5**	**215.3**	**214.9**	**-0.4**
Average score if all SD excluded were assigned a score two standard deviations below the mean for SD tested	**201.7**	**207.9**	**6.2**	**211.5**	**210.8**	**-0.7**

The first thing to note is the simple fact that the last two rows differ, and ,of course, that neither matches the results for students actually tested. If the intended inference from

[5] The standard deviations of scores for tested students with disabilities were as follows: For New York City, 31.55 (2002) and 34.65 (2003), and for the United States, 38.21 (2002) and 40.17 (2003).

NAEP results is to the entire student population, then exclusions contribute to uncertainty. The second point is that even where the exclusion rate is constant, as with the five percent excluded nationally in 2002 and again in 2003, exclusions may affect score comparisons. The estimated 2002-to-2003 change for the Nation under the scenario in the bottom row is over twice as large as the change observed, although in this case still not statistically significant. The third point is that when exclusion rates are <u>not</u> constant, the effects of exclusions on comparisons can be dramatic. Comparing the second and last rows of the table for New York City, it is seen that adjusting for the 5% exclusions in 2002 depressed the score estimate by 2.9 points, whereas adjusting for the 2% exclusions in 2003 depressed the score estimate by just 1.3 points. Consequently, the estimated change from 2002 to 2003 was much larger. It is reasonable to conclude from Table 1 that New York City's greater success in 2003 in testing as many students as possible slightly depressed the estimated fourth-grade improvement over the previous year. The 5.3 point change in the bottom line would probably be statistically significant.

The final point to be made using Table 1 is that effects of exclusions must be interpreted relative to the background uncertainty summarized by the standard error. Consider the second row from the top versus the second row from the bottom. For New York City in 2002, the difference is 206.2 - 204.8 = 1.4 points. For the United States in 2002, the difference is 216.8 - 215.3 = 1.5 points, nearly the same. But the New York City value is barely over half a standard error (1.4 is just over half of 2.6) whereas the United States value is three times the standard error (1.5 is three times 0.5). That means that the New York City change is quite minor relative to the overall uncertainty as to the city's performance, whereas the change for the United States is quite substantial. <u>This point is increasingly important because over time, NAEP is becoming more and more precise</u>. Several years ago, with the change to contractor administration for State NAEP, it became possible to merge the state and national NAEP samples, resulting in dramatic improvements in accuracy. "No Child Left Behind" mandates for NAEP participation have brought all fifty states into state NAEP for the first time, further increasing sample sizes, and have greatly reduced nonparticipation at the school level. Finally, the initiation of the Trial Urban District Assessment has also increased sample sizes in many states and for the Nation. Standard errors of NAEP scores are much smaller now than five years ago, and so the effects of exclusions loom much larger relative to other sources of error.

Table 1 illustrates effects of exclusion on trends over time, but the same considerations apply to cross-sectional comparisons among jurisdictions. Table 2 presents information about the ten districts that participated in the 2003 Trial Urban District Assessment, again using fourth-grade reading. The first column identifies the district and the second column tells what percentage of all sampled students were *identified* as students with disabilities and/or LEP students. Some identified students are tested without accommodations, some are tested with accommodations, and some are excluded. The wide variation in the percents of students identified helps to explain the wide variation shown in the third column, which gives the percents of students excluded. In Atlanta, for example, about one in fifty sampled students were excluded for reasons of SD and/or LEP designations. In Houston, that number was close to one in four. Columns four and five further break down the exclusions according to SD versus LEP. (Students designated LEP are

Table 2. Students with disabilities and limited-English-proficient students identified and excluded, as a percentage of all fourth-grade students in public schools, by urban district, 2003

District	Percent SD &/or LEP Identified	Percent SD &/or LEP Excluded	Percent SD Excluded	Percent LEP Excluded	Mean Reading Score	Mean for incl. SD	Std Dev for incl. SD	Mean for incl. LEP	Std Dev for incl. LEP
Atlanta	9	2	2	1	240	180	44	----	----
Boston	33	9	4	6	252	181	32	192	32
Charlotte	21	5	4	3	262	191	35	190	30
Chicago	31	9	6	6	248	163	42	176	35
Cleveland	18	12	11	2	240	161	29	----	----
District of Columbia	18	6	5	1	239	148	43	174	40
Houston	42	24	9	20	246	183	32	186	30
Los Angeles	59	6	3	5	234	167	38	183	33
New York City	21	6	2	5	252	181	35	183	31
San Diego	42	5	3	4	250	185	39	186	34

excluded if they have received instruction in English for less than three years and are judged by school staff to be incapable of participating in the assessment in English.) Because some students are both SD and LEP, the sum of the values in these two columns may exceed the value in column three. The rightmost five columns give the mean fourth-grade reading score for all students, followed by the mean and standard deviation of reading scores for tested students with disabilities and for tested limited English proficient students. Note that in Atlanta and Cleveland, the numbers of LEP students were too small to produce estimates sufficiently reliable for reporting.

Table 2 reveals some of the challenges in standardizing procedures around exclusions and accommodations. The wide variation in proportions identified and excluded may reflect different state policies, but primarily reflect differences in student populations. These differences are also shown by comparisons of the means for students with disabilities versus LEP students. In some districts, these means are virtually the same, and in other districts, they differ markedly.

Table 3. 2003 TUDA, fourth-grade reading, alternative imputations for excluded students' scores

City	Mean Reading Score	Rank	Estimate Using SD &/or LEP Means	Rank	Estimate Using 2 Std Dev Below	Rank
Atlanta	240	8	238.34	6	236.59	5
Boston	252	2	246.60	4	240.90	4
Charlotte	262	1	258.35	1	255.06	1
Chicago	248	5	241.26	5	234.36	6
Cleveland	240	7	230.85	10	223.85	9
District of Columbia	239	9	233.51	7	228.41	7
Houston	246	6	231.32	8	216.48	10
Los Angeles	234	10	230.87	9	226.65	8
New York City	252	3	247.64	2	243.80	2
San Diego	250	4	246.78	3	243.16	3

Table 3 presents some hypothetical illustrations of the possible effects of exclusions on district comparisons. It bears repeating that there is no entirely satisfactory way to "adjust" for the effects of exclusions, but the calculations in Table 3 may give some rough sense of the magnitude of effects of exclusions. The same scenarios are used here as were used in Table 1. The second column repeats the mean fourth-grade reading scores shown in Table 2, and the third column shows the district rankings according to those scores. The fourth and fifth columns show means and ranks if excluded SD and LEP students were included and earned average scores equal to the average scores for tested students in those respective groups, and the final two columns show means and ranks if excluded students were included and earned average scores two standard deviations below the mean.[6] Under these alternative scenarios, ranks for some districts change substantially.

In addition to comparisons over time and comparisons among states or districts, differential exclusion rates can affect comparisons among subgroups. Achievement gaps among White non-Hispanic, Black, and Hispanic students have long been of concern, and attention to gaps and gap closing has been heightened by the No Child Left Behind Act of 2001. Exclusion rates vary among subgroups to an even greater extent than among states or districts. In the fourth-grade reading assessment in 2003, percentages of tested students classified by the school as having a disability (IEP) ranged from 6 percent for Asian/Pacific Islander to 13 percent for American Indian. Percentages for White, Black, and Hispanic students were 10, 10, and 9 percent, respectively. It is plausible that the percentages of students excluded were roughly proportional to the percentages identified among those tested. For the same assessment, percentages of tested students classified by the school as LEP ranged from 1 percent for White students and for Black students to 37 percent for Hispanic students. Figures for Asian/Pacific Islander and American Indian students were 21 percent and 15 percent, respectively. Again, LEP exclusions would be expected to be roughly proportional to the percentages identified among those tested.

Effects of Exclusions on Reliability

Reliability refers to the accuracy or reproducibility of scores for individuals or, in the case of NAEP, of means, etc. for student populations. In the context of NAEP, reliability is captured by the standard errors of NAEP statistics.[7] Exclusions affect reliability in at least two ways. First, if some students are excluded from participating, the sample size is diminished. Because sample size is related to precision, higher exclusion rates will result in slightly larger standard errors. This effect is easy to quantify, and also easy to overcome, albeit at some cost. A larger sample may be drawn initially to yield the target sample size after exclusions.

The second effect of exclusions on reliability is potentially more serious, and much more difficult to quantify. Decisions about including versus excluding individual students involve human judgments, and any such judgments introduce a degree of uncertainty. Inclusion/exclusion decisions for SD and LEP students depend on judgments captured in federal, state, district, and school policies and practices concerning student referral processes, the creation of IEPs, bilingual education, and student testing. Uncertainty due to judgments at the local school level could be offset by increasing sample size, although the amount of increase required would be difficult to determine.

[6] For purposes of these hypothetical calculations, percent SD excluded and percent LEP excluded were adjusted proportionally so that they summed to the percent SD &/or LEP excluded. For Atlanta and Cleveland, means and standard deviations for SD students were substituted for the corresponding values for LEP students.

[7] Reliability may also be expressed using a coefficient with a value between zero and one, representing the correlation between actual or hypothetical replications of the measurement procedure, or the proportion of score variance not attributable to variation across replications of the measurement procedure. Whether expressed using reliability coefficients or standard errors, the essential concern is accuracy, replicability, stability, or precision of individual or group scores.

Of more concern are the effects of judgments at the district or state levels. Statutes, laws, policies, regulations, and definitions at the state level have the potential to affect exclusion decisions for all students in the state. If two states differ in their exclusion practices, the accuracy of comparisons between those two states will be compromised in a manner that cannot be offset by increasing sample sizes. In fact, if sample sizes within the two states are increased, their NAEP standard errors will become smaller, and the distortions due to state-level differences in exclusion practices will become more and more salient relative to the background uncertainty of the assessment. Likewise, if exclusion practices within a state change over time, the accuracy of trend lines for that state will be compromised.[8]

The effects of exclusions on reliability can be minimized first, by minimizing exclusions; second, by establishing exclusion criteria that are as clear and objective as possible and working to assure that those criteria are adhered to; and third, by making practices and criteria across states as uniform as possible.

Effects of Exclusions on Validity

Validity is more complex than reliability. The term is used in many ways, but may be defined here as the soundness and accuracy of inferences from NAEP findings. Any exclusions affect validity because they distort the picture of student achievement portrayed by NAEP reports. In this section, these distortions, their consequences, and their remedies are discussed under three subheadings. The first addresses the general effects of any exclusions. The second and third address the effects of differential exclusion rates arising from different causes. The exclusion rate for a state, a district, or a demographic group is a function of two factors. One factor comprises the formal criteria for exclusion and the way those criteria are actually applied. The other factor is the proportion of students who truly meet those criteria. Thus, variations in exclusion rates may arise due to variations in exclusion criteria (and the ways they are implemented) and/or variations in the proportions of students meeting some ideal uniform, error-free criteria for exclusion due to disability, limited English proficiency, or both. The second and third subheadings address the effects of these two sources of variation, respectively.

- Effects of exclusions per se

It is NAEP's goal to be as inclusive as possible of all public-school students enrolled at each grade assessed.[9] The overwhelming majority of reports of NAEP findings in the media make no mention of exclusions, simply referring to the performance of fourth-

[8] Appendix A of the NAEP 2002 Reading Report Card includes analyses of the potential effects of exclusion rates on assessment results. These analyses were prompted in part by the finding of a modest correlation between changes in exclusion rates from 1998 to 2002 and changes in states' average test scores. The effects of different state policies on exclusion rates might be framed either as a problem of reliability or as a problem of validity, but for present purposes there is no need to pursue that discussion.

[9] Obviously, the same considerations apply to NAEP private-school samples, samples defined by age rather than grade, etc.

grade students in California, for example. Even in NAEP reports, although appendices explain exclusions, most tables and figures and the accompanying text omit any caveats or reminders about exclusions. Thus, it is abundantly clear that the intended inferences from NAEP are primarily to entire student populations. Any exclusions, from whatever cause, bias estimates of achievement distributions and thereby distort the accuracy of those inferences. Because excluded students would, on average, score lower than students tested, the effect of exclusions is to bias apparent achievement upward.

The illustrations earlier in this paper focused on average scores, but the effects of exclusions are also seen in descriptions based on achievement levels. In 2003, for example, NAEP reported that 30 percent of fourth grade students were at or above Proficient in reading. In that sample, 22 percent of fourth graders were identified as having a disability or being limited English proficient, and 6 percent of all fourth graders sampled were excluded. It is impossible to know what proportion of those excluded students were proficient, but if none of them were, the actual national percent at or above proficient would be 28 percent, not 30 percent.[10] The actual bias is somewhat smaller, but to the nearest whole number, 28 percent is probably correct.[11]

The consequences of such systematic distortions are probably small, so long as (1) overall exclusion rates are low and (2) exclusion rates are uniform. It is difficult to imagine a consequential policy decision that would turn on the difference between 28 percent and 30 percent proficient in the above example. As shown in Table 1, however, exclusions can affect the magnitude of estimated differences among groups or of estimated changes over time, even if the exclusion rates are uniform. Moreover, as discussed below, nonuniform exclusion rates have more serious consequences.

Remedies for this general effect of exclusions are clear. First, efforts to minimize exclusions should continue. The higher the proportion of exclusions, the greater the bias and the greater the uncertainty about the true performance level of the population. Second, NAEP users should perhaps be reminded more often that the students tested do not represent the entire population. Finally, research should continue on the utility of imputation models that might be used to adjust for effects of exclusions.

- Effects of variation in formal exclusion criteria and the ways they are implemented

Nearly all inferences from NAEP involve comparisons. The 30 percent (or 28 percent) proficiency rate for the Nation's fourth graders in 2003 is interpreted by noting that the rate is unchanged since 2002 or 2000, but that it is significantly higher than the rate in 1998. The actual difference between 1998 and 2003 is small, however--just 2 percent--

[10] The figure 28 percent is a weighted average of a 30 percent (actually, 29.67 percent) rate for the 94 percent of students tested, and an assumed 0 percent rate for the 6 percent of students excluded.
0.94 x 29.67 + 0.06 x 0 = 27.89, which rounds to 28 percent.

[11] Among students with disabilities who were tested, only 9 percent were at or above proficient, and among limited English proficient students who were tested, only 7 percent were at or above proficient. Thus, the actual proportion proficient among excluded students is almost surely less than 8 percent. Imputing 8 percent proficient among excluded students gives an upper bound for at-or-above proficient of 28.36 percent.

and the significance test was close to the threshold between "significant" and "not significant."[12] The percent of students excluded changed between 1998 and 2003, from 7 percent to 6 percent. In this instance, the direction of the change in exclusion rates would have served to reduce the apparent gain. The gain in percent proficient was declared significant despite being slightly understated. Cases will inevitably arise, however, where changes in exclusion rates will affect conclusions as to whether changes over time were significant or not. The same holds for contrasts among jurisdictions and for gaps among demographic groups.

When NAEP performance comparisons are reported, whether over time, among states or districts, or among subgroups (as with "performance gaps"), the intended inference is that the observed difference accurately mirrors the difference in performance for the populations compared. More remote inferences may be also be drawn--That achievement gaps reflect socioeconomic differences, for example--but the first step is always an inference from what the test data show to what the students actually know and can do.

The consequences of differential exclusion policies are serious. If such policies vary for the two time points, groups, or jurisdictions compared, then that first, fundamental inference is compromised. The observed change, contrast, or performance gap in fact represents some mixture of differences in actual student achievement distributions and differences in decision rules determining whom to test. The comparison of performance for tested groups is thus "apples to oranges." As noted earlier, the increasing accuracy of NAEP statistics has made even small distortions more important than they once were.

At one level, the formal exclusion criteria for NAEP are uniform, but those uniform criteria make reference to local policies surrounding IEPs, and allow for substantial local discretion with regard to the testing of LEP students. An analysis of variations in policies and practices is outside the scope of this paper, but it seems clear from the observed variation in exclusion rates across states that many students excluded in one state would have been included if they resided elsewhere, and conversely. Among the states participating in the 2002 fourth-grade reading assessment, SD and/or LEP exclusion rates ranged from 3 percent to 12 percent. Some high rates (e.g., Texas, 11 percent) reflect demographic differences, especially the prevalence of LEP students. Others (e.g., North Carolina, 12 percent) may be due at least in part to variation in states' implementation of SD and LEP identification and exclusion policies. Overall, for 2002 fourth-grade reading, the correlation across states between the percent of all students identified as SD and/or LEP and the percent of all students excluded as SD and/or LEP was only .38. If policies were uniform, a higher correlation would be expected. For the 36 states that participated in both the 1998 and the 2002 fourth-grade reading assessment, the correlation between the change in exclusion rate from 1998 to 2002 and the change in achievement from 1998 to 2002 was .57. This indicates that a substantial proportion of variation in states' achievement score changes can be accounted for by changes in their rates of exclusion.

[12] The NAEP Data Tool shows percent at-or-above proficient in 1998 as 27.51 percent (standard error 1.01) and percent at-or-above proficient in 2003 as 29.67 percent (standard error .30). The difference is statistically significant at a p value of .0434, just slightly below the .05 threshold.

Remedies for differential policies are obvious--The actual implementation of exclusion policies should be as uniform as possible over time, across jurisdictions, and, of course, across demographic groups. The fundamental purpose of this background paper is to support and encourage deliberations leading to such uniformity.

- Effects of variation in underlying rates of disability and limited English proficiency

Surely, the major reason for variation in exclusion rates is variation in the proportions of students who truly meet the NAEP criteria. The incidence of many cognitive disabilities is negatively related to socioeconomic status (i.e., children in poverty are more likely to have special needs), although the rate of diagnosis of learning disabilities may be positively related to socioeconomic status (i.e., children in more affluent households are more likely to receive the help they need). In short, SD incidence may be related to socioeconomic variables. In addition, there may be some migration of families whose children have certain special needs to states with exemplary services for such students. Thus, identification and exclusion rates for students with disabilities would be expected to vary across states.

There is enormous state-to-state variation in the numbers and demographics of students with limited English proficiency. Although most LEP students in the United States are Hispanic, California has substantial numbers of Asian LEP students, as well. Some states have virtually no LEP students. In California, LEP students make up a quarter of the student population.

There are consequences to population differences in the incidence of SD and/or LEP students meeting NAEP exclusion criteria. Even though such differences are real, they are still problematic. Estimates of the White-Hispanic gap, for example, are distorted by the differential exclusion of Hispanic versus White students due to limited English proficiency. Likewise, comparisons between states with high versus low exclusion rates will be distorted to some degree. Differences in student demographics from state to state may partially explain the low correlation between percent of SD and/or LEP students identified and percent of SD and/or LEP students excluded. The typical LEP student in Texas may be different from the typical LEP student in California, and so the probability that an identified student meets exclusion criteria may vary from one state to another.

There is no real remedy for demographic variation. The effects of such differences on inferences from NAEP can be minimized by pursuing the remedies already described: minimizing exclusions overall, clearly explaining the effects of exclusions in reports of NAEP findings, striving toward uniformity in exclusion policies and practices, and continuing to investigate the feasibility of statistical adjustments to reduce the bias from exclusions.

Conclusion

As "The Nation's Report Card," NAEP is designed to represent the performance of all students at selected grade levels. Exclusions of SD and/or LEP students introduce bias in all NAEP statistics, and affect conclusions as to whether changes over time, contrasts among jurisdictions, or gaps among subgroups are statistically significant. Even if exclusion rates are the same for two populations compared, they can affect the comparison. When exclusion rates differ, effects can be substantial.

The size and importance of exclusion effects must be judged relative to the background uncertainty in NAEP statistics, as represented by their standard errors. Recent changes in NAEP have greatly increased its accuracy (reduced standard errors) and for that reason, exclusion effects are increasingly important.

For the most part, effects of exclusions on reliability can be offset by increasing sample sizes. Effects of exclusions on validity are more problematical. It is important that NAEP continue to keep exclusions to a minimum, and that efforts be made to work toward more uniform policies and practices for determining which students should be excluded from NAEP and which should be tested.