A Methodological Notes

This appendix describes in more detail some of the methods used in this report. It covers technical aspects of (a) the placement of state achievement standards on the NAEP scale, and (b) the construction of a population achievement profiles based on school level averages. Finally, the estimation of the achievement of NAEP excluded students is discussed briefly; further details regarding this methodology as applied to reading assessments can be found in McLaughlin (2003).

Estimating the Placement of State Achievement Standards on the NAEP Scale

If an achievement standard can be operationalized as a cutpoint on the NAEP scale, it is straightforward to estimate the percentage of the students in a state who meet that standard from the NAEP data. One compares each *plausible value* on the achievement scale assigned to a NAEP student, based on his or her responses to the test items, to the cutpoint of the standard. If it is greater than the cutpoint, the student's *weight* (the number of students in the population he or she represents) is added to the count of those meeting the standard; otherwise it is added to the count of those not meeting the standard.

If we had both the NAEP data for a state and the percentage of students in the state who met a NAEP standard, sorting all the plausible values in ascending order and determining which one just corresponds to the percent meeting the standard would be a straightforward task. For example, if the percent meeting the standard were given as 25, we would count down from the top of the order, adding up the weights until we reached 25 percent of the total weight for the state. This would not be exact, because there is some space between every pair of plausible values in the database, but with typically more than 2,000 NAEP participants in a state, we would expect it to be very close.

In this equipercentile mapping method, the standard error is an estimate of how far our estimate can be wrong, on average. The standard error is clearly related to the number of NAEP participants in the state.

Next, suppose that the percent meeting the standard is for the state's own assessment of achievement, not for NAEP's standard. We could carry out the same procedure to

find an estimate of the NAEP scale value corresponding to the state's standard; that is, the *cutpoint* for the state standard. Again, its standard error would depend on how large the NAEP sample of students is.

The method of obtaining *equipercentile equivalents* involves the following steps:

a. obtain for each school in the NAEP sample the proportion of students in that school who meet the state performance standard on the state's test;

b. estimate the state proportion of students meeting the standard on the state test by weighting the proportions (from *step a*) for the NAEP schools, using NAEP weights;

c. estimate the weighted distribution of scores on the NAEP assessment for the state as a whole, based on the NAEP sample of schools and students within schools, and

d. find the point on the NAEP scale at which the estimated proportion of students in the state scoring above that point (using the distribution obtained in *step c*) equals the proportion of students in the state meeting the state's own performance standard (obtained in *step b*).

Operationally, the reported percentage meeting the state's standard in each NAEP school s, $p_s^{[STATE]}$, is used to compute a state percentage meeting the state's standards, using the NAEP school weights, w_s. For each school, w_s is the sum of the student weights, w_{is}, for the students selected for NAEP in that school.[1] For each of the five sets of NAEP plausible values, v=1, 2, 3, 4, 5, we solve the following equation for c, the point on the NAEP scale corresponding to the percentage meeting the state's standard:

$$p^{[STATE]} = \sum_{is} w_{is} p_s^{[STATE]} \Big/ \sum_{is} w_{is} = \sum_{is} w_{is} \partial_{isv}^{[NAEP]}(c) \Big/ \sum_{is} w_{is}$$

where the sum is over students in schools participating in NAEP, and $\partial_{isv}^{[NAEP]}(c)$ is equal to 1 if the *v-th* plausible value for student i in school s, y_{isv}, is greater than or equal to c. The five values of c obtained for the five sets of plausible values are averaged to produce the NAEP threshold corresponding to the state standard.

Specifically, each of the five parallel sets of NAEP plausible values (in the combined set of NAEP schools with matching state data) is sorted in increasing order. Then, for each plausible value in set v, y_v, the percentage of the NAEP student distribution that is greater than or equal to y_v, $p_v^{[NAEP]}(y_v)$, is computed. The two values of $p_v^{[NAEP]}(y_v)$ closest to $p^{[STATE]}$ above and below are identified, $p_{Uv}^{[NAEP]}(y_{Uv})$ and $p_{Lv}^{[NAEP]}(y_{Lv})$; and a point solution for c_v is identified by linear interpolation between y_{Uv} and y_{Lv}.

1. To ensure that NAEP and state assessments are fairly matched, NAEP schools which are missing state assessment scores (i.e., small schools, typically representing approximately four percent of the students in a state) are excluded from this process. Even if the small excluded schools are higher or lower performing than included schools, that should introduce no substantial bias in the estimation process, unless their high or lower scoring was specific to NAEP or specific to the state assessment.

Variation in results over the five sets of plausible values is a component of the standard error of the estimate, and the average of the five results is the reported mapping of the standard onto the NAEP scale.

The problem with this simple method is that it could be applied to any percentage, not just the percent meeting the state's achievement standard. Finding which NAEP scale score corresponds to the percent of the students, say, *living in large cities* would yield a value that is meaningless. A method is needed for testing the assumption that the percent we are given really corresponds to achievement on the NAEP scale.

The method we use to test the assumption is based on the fact that we are given the percent meeting the standard for each school participating in NAEP (and NAEP tests a random, representative sample of grade 4 or grade 8 students in each participating school). If the percentage we obtain from the state assessment for each school corresponds to achievement on the NAEP scale in that school, then applying the cutpoint, c, we estimated for the whole state to the NAEP plausible values in that school should yield an estimate of the percent meeting the standard in that school, based on NAEP, which matches the reported percent from the state assessment:

$$p_s^{[STATE]} = p_s^{[NAEP]}(c).$$

Of course, our estimated percentage meeting the state standard will not be exactly the same as the reported percent, because (a) NAEP only tests 20 to 25 students in the school, and (b) tests are not perfectly reliable. Moreover, in some states, we are given a grade 5 score for the state standard; in such cases, for the mapping method to be valid, we must assume that, on average across the state, the same percent of fourth graders would meet a grade 4 achievement standard as the percent of fifth graders who met the grade 5 standard. Of course, that would mean that our estimate for each school would have greater error—not only do some students learn more between fourth grade and fifth grade than others, but each cohort of students is different from the preceding grade's, in both random and systematic ways.

We need to have an estimate of how much error to expect in guessing each school's percent meeting the state's standard, to which we can compare the actual error. For this, we estimate the size of error if the state's standard were actually parallel to NAEP, and check whether the observed error is more similar to that or to the size of error we would expect if the reported percentages for the schools were unrelated to NAEP. If the error is sufficiently large, that calls the accuracy of the estimated standard into question.

Test criterion for the validity of the method

The test criterion is based on an evaluation of the discrepancies between (a) individual schools' reported percentages of students meeting a state standard and (b) the percentages of the NAEP achievement distribution that are greater than the NAEP cutpoint estimated for the state to be equivalent to that state standard. The method of estimation ensures that, on the average, these percentages agree, but there

is no assurance that they match for each school. To the extent that NAEP and the state assessment are parallel assessments, the percentages should agree for each school, but if NAEP and the state assessment are not correlated, then the linkage will not be able to reproduce the individual school results.

Failure of school percentages to match may also be due to student sampling variation, so the matching criterion must be a comparison of the extent of mismatch with the expectation based on random variation. To derive a heuristic criterion, we assume linear models for the school's reported percentage meeting standards on the state test, p_s, and the corresponding estimated percentage for school s, $\hat{p}_s$. The estimated percentage $\hat{p}_s$ is obtained by applying the linkage to the NAEP plausible values in the school.

$$p_s = \pi + \lambda_s + \delta_s + \gamma_s$$
$$\hat{p}_s = \pi + \lambda'_s + \delta'_s + \gamma_s$$

where π is the overall mean, λ and λ' are separate true between-school variations unique to the two assessments, δ and δ' are random sampling variations (due to the finiteness of the sample in each school), and γ is the common between-school variation between the reported and estimated percentages. We hope γ is large and λ and λ' are small. If λ and λ' are zero then the school-level scores would be reproducible perfectly except for random sampling and measurement error. Although linear models are clearly an oversimplification, they provide a way of distinguishing mappings of standards that are based on credible linkages from other mappings.

In terms of this model, the critical question for the validity of a mapping is whether the variation in γ, the achievement common to both assessments, as measured by $\sigma^2(\gamma)$, is large relative to variation in λ and λ', the achievement that is different between the assessments, as measured by $\sigma^2(\lambda)$ and $\sigma^2(\lambda')$. Because the linkage is constructed to match the variances of p_s and $\hat{p}_s$, the size of the measurement error variance, $\sigma^2(\delta)$ and $\sigma^2(\delta')$, does not affect the validity of the standard mapping, although it would affect the reproducibility of the school-level percentages.

The relative error criterion is the ratio of variance estimates:

$$k = [\sigma^2(\gamma) + (\sigma^2(\lambda) + \sigma^2(\lambda'))/2\ ]/\sigma^2(\gamma)$$

The value of k is equal to 1 when there is no unique variance and to 2 when the common variance is the same as the average of the unique variances. Values larger than 1.5 indicate that the average unique variation is more than half as large as the common variation, which raises concern about the validity of the mapping.

We can estimate variance components from the variances of the sum and difference between the observed and estimated school-level percentages meeting a standard.

$$\sigma^2(p + \hat{p}) = \sigma^2(\lambda) + \sigma^2(\delta) + \sigma^2(\lambda') + \sigma^2(\delta') + 4\sigma^2(\gamma), \text{ and}$$

$$\sigma^2(p - \hat{p}) = \sigma^2(\lambda) + \sigma^2(\delta) + \sigma^2(\lambda') + \sigma^2(\delta').$$

By subtraction,

$$\sigma^2(\gamma) = (\sigma^2(p + \hat{p}) - \sigma^2(p - \hat{p}))/4.$$

To estimate $\sigma^2(\lambda)$ and $\sigma^2(\lambda')$, we compute $\sigma^2(p - \hat{p})$ for a case in which we know that there is no unique variation: using two of the five NAEP plausible value sets for the observed percentages and the other three for the estimated percentages. In this case,

$$\sigma^2(p_{NAEP} - \hat{p}_{NAEP}) = \sigma^2(\delta) + \sigma^2(\delta').$$

Substituting this in the equation for the variance of the differences, and rearranging terms,

$$\sigma^2(\lambda) + \sigma^2(\lambda') = \sigma^2(p - \hat{p}) - \sigma^2(p_{NAEP} - \hat{p}_{NAEP})$$

Substituting these estimates into the equation for k, we have

$$k = [\sigma^2(p + \hat{p}) - \sigma^2(p - \hat{p}) + 2\sigma^2(p - \hat{p}) - 2\sigma^2(p_{NAEP} - \hat{p}_{NAEP})] / (\sigma^2(p + \hat{p}) - \sigma^2(p - \hat{p}))$$

$$k = 1 + 2\,[\sigma^2(p - \hat{p}) - \sigma^2(p_{NAEP} - \hat{p}_{NAEP})/(\sigma^2(p + \hat{p}) - \sigma^2(p - \hat{p}))$$

That is, k is greater than 1 to the extent that the differences between observed and estimated school-level percentages are greater than the differences would be if both assessments were NAEP. [2]

The median values of k for primary grades 4 and 8 mathematics standards across states in 2003 were 1.241 and 1.171, corresponding to median values of 0.115 and 0.090 for $\sigma(p - \hat{p})$, 0.041 and 0.040 for $\sigma(p_{NAEP} - \hat{p}_{NAEP})$, and 0.216 and 0.197 for

2. The fact that the simulations are based on subsets of the NAEP data might lead to a slight over-estimate of $\sigma^2(\delta) + \sigma^2(\delta')$ because the distributions would not allow as fine-grained estimates of percentages achieving the standard. That over-estimate of random error in the linkage would, in turn, slightly reduce the estimate of k. In future work, one alternative to eliminate that effect might be to create a *parallel* NAEP by randomly imputing five additional plausible values for each participant, based on the mean and standard deviation of the five original plausible values. The result might increase the relative error measures slightly because they might reduce the term subtracted from the numerator of the formula for k.

$\sigma(p + \hat{p}) / \sqrt{2}$, when p is measured on a [0,1] scale. A value of 1.5 for k corresponds to a common variance equal to the sum of the unique reliable variances in the observed and estimated percentages meeting the standards.

Setting the criterion for the validity of this application of the equipercentile mapping method at k = 1.5 (signifying equal amounts of common and unique variation) is arbitrary but plausible. Clearly, it should not be taken as an absolute inference of validity—two assessments, one with a relative criterion ratio of 1.6 and the other with 1.4, have similar validity. Setting a criterion serves to call attention to the cases in which one should consider a limitation on the validity of the mapping as an explanation for otherwise unexplainable results. While estimates of standards with greater relative error due to differences in measures are not, thereby, invalidated, any inferences based on them require additional evidence. For example, a finding of differences in trend measurement between NAEP and a state assessment when the standard mapping has large relative error may be explainable in terms of unspecifiable differences between the assessments, ruling out further comparison. Nevertheless, because the relative error criterion is arbitrary, results for all states are included in the report, irrespective of the relative error of the mapping of the standards.

Notes

With the relative error criterion we assessed the extent to which the error of the estimate is larger than it would be if NAEP and the state assessment were testing exactly the same underlying trait; in other words, by evaluating the accuracy with which each school's reported percentage of students meeting a state standard can be reproduced by applying the linkage to NAEP performance in that school. The method discussed here ensures that, on average, these percentages match, but there is no assurance that they match for each school. To the extent that NAEP and the state assessment are parallel assessments, the percentages should agree for each school, but if NAEP and the state assessment are not correlated, then the mapping will not be able to reproduce the individual school results. One difficult step in the validation process was estimating the amount of error to expect in reproducing the state-reported percentages for schools that could be due to random measurement and sampling error and not due to differences in the underlying traits being measured. For this purpose, we estimated the amount of error that would exist if both tests were NAEP. We used the distribution based on two plausible value sets to simulate the observed percent achieving the standard in each school and the distribution of the other three plausible value sets to simulate the estimated percent achieving the standard in the same school. The standard was the NAEP value determined (based on the entire state's NAEP sample) to provide the best estimate of the state's standard. Given the standard (a value on the NAEP scale), the percents achieving the standard are computed solely from the distribution of plausible values in the school. As an example, suppose the estimated standard is 225. For a school with 25 NAEP participants, there would be a distribution of 50 plausible values (two for each student) in the school for the simulated observed percent and 75 (three for each student) for the simulated estimated percent. The 50 plausible values for a school

represent random draws from the population of students in the state who (a) might have been in (similar) schools selected to participate in NAEP and (b) might have been selected to respond to one of the booklets of NAEP items. That distribution should be the same, except for random error, whether it is based on two, three, or five sets of plausible values.

Constructing a Population Achievement Profile Based on School-level Averages

For this report, individual scores on state assessments were not available. The comparisons in the report are based on school-level state assessment statistics and corresponding school-level summary statistics for NAEP. These school-level statistics include demographic breakdowns; that is, summary statistics for Black students in each school, Hispanic students in each school, and students eligible for free/reduced price lunch in each school. These are used in comparing NAEP and state assessment measurement of achievement gaps.

As defined in this report, a *population profile* of achievement is a percentile graph showing the distribution of achievement, from the lowest-performing students in a group at the left to the highest-performing students at the right. Concretely, one can imagine the students in a state lined up along the *x*-axis (i.e., the horizontal axis of a graph), sorted from left to right in order of increasing achievement scores, with each student's achievement score marked above him/her.[3]

When achievement scores are only available as school averages, or school averages for a particular category of students, the procedure, and the interpretation, is slightly different. Imagine the state's students in a demographic group lined up along the *x*-axis, sorted in order of average achievement of their group in their school (e.g., the percentage of students in the group who meet an achievement standard). Students in a school would be clustered together with others of their group in the same school. Each school's *width* on the *x*-axis would represent the *weight* of its students in representing the state population for the group. Thus, a school with many students in the demographic group would take up more space on the *x*-axis.

The population profile would then refer to the average performance in a school for the particular demographic group. The interpretation is similar to population profiles based on individual data, but there are fewer extremely high and extremely low scores. Gaps have the same average size because the achievement of each member of each demographic group is represented equally in the individually based and the school-level based profiles.[4] It is important to note that when we refer to school-level data, we are referring to aggregate achievement statistics for separate demographic groups in each school.

Because each school is weighted by the number of students in the demographic group it represents, we can still picture the population achievement profile as lining up the students in a state. They would be sorted from left to right in increasing order of their

3. See figure 1 in the text for an example of a population profile.
4. The exception to this is that due to suppression of small sample data in state assessment publications. As a result, students in schools with very small representations of a demographic group are underrepresented in school-level aggregates.

school's average achievement for their demographic group, with that average marked above them.

The procedure for computing a population profile based on school-level data can be described mathematically as follows. We start with a set of schools, j, in state i with N_i schools, with average achievement y_{igj}, for group g, whose weight in computing a state average for group g is w_{igj}. Sorting the schools on y_{igj}, creates a new subscript for schools, k, such that $y_{igk} \le y_{ig(k+1)}$.

The sequence of histograms of height y_{igk}, width w_{igk}, for k=1,…, N_i, forms a continuous distribution, which can be partitioned into one hundred equal intervals, c = 1,…, 100, or percentiles. The achievement measure y_{igc}, for interval c in demographic group g in state i, is given by

$$y_{igc} = \left(\sum_{k = A_{igc}}^{B_{igc}} \bar{w}_{igk} \cdot y_{igk} \right) \Big/ \left(\sum_{k = A_{igc}}^{B_{igc}} \bar{w}_{igk} \right)$$

where the A_{igc}, B_{igc}, and $\bar{w}_{igk}$ values are defined as follows:

$$\sum_{k = 1}^{A_{igc} - 1} w_{igk} \le \frac{(c-1)W}{100} \le \sum_{k = 1}^{A_{igc}} w_{igk} \text{ and}$$

$$\sum_{k = 1}^{B_{igc} - 1} w_{igk} \le \frac{(c)W}{100} \le \sum_{k = 1}^{B_{igc}} w_{igk}$$

where W is the total weight.

For $A_{igc} + 1 \le k \le B_{igc} - 1$,

$$\bar{w}_{igk} = w_{igk};$$

$$\bar{w}_{A igc} = \left(\sum_{k = 1}^{A_{igc}} w_{igk} - \frac{(c-1)W}{100} \right), \text{ and}$$

$$\bar{w}_{B igk} = \frac{(c)W}{100} - \left(\sum_{k = 1}^{B_{igc} - 1} w_{igk} \cdot w_{B_{igc}} \right).$$

It is important to note that the ordering of the schools for one group (e.g., Black students) may not be the same as the ordering of the same schools for another group (e.g., White students). Therefore, the gap between two population achievement profiles is not merely the within-school achievement gap; it combines both within-school and between-school achievement gaps to produce an overall achievement gap.

Standard errors can be computed for the individual y_{igc} by standard NAEP estimation methodology, computing a profile for each set of plausible values and for each set of replicate weights. However, in this report, we combined the percentiles into six groupings: the lowest and highest quartiles, the middle 50 percent, the lower and upper halves, and the entire range. The comparison of achievement between two groups for the entire range is mathematically equivalent to the average gap in the selected achievement measure.

Estimating the Achievement of NAEP Excluded Students

Since 1998, there has been concern that increasing or decreasing rates of exclusion of NAEP students from the sample might affect the sizes of gains reported by NAEP (e.g., Forgione, 1999; McLaughlin 2000, 2001, 2003). A method for imputing plausible values for the achievement of the excluded students based on ratings of their achievement by school staff has been applied to produce *full-population estimates*. The following description of the method is excerpted from a report to NCES on the application of the method to the estimation of reading gains between 1998 and 2002 (McLaughlin, 2003). The same method was used to produce full population estimates for mathematics in 2000 and 2003.

The method is made possible by the NAEP SD/LEP questionnaire, a descriptive survey filled out for each student with disability or English language learner selected to participate in NAEP - whether or not the student actually participates in NAEP or is excluded on the grounds that NAEP testing would be inappropriate for the student. The basic assumption of the method is that excluded students in each state with a particular profile of demographic characteristics and information on the SD/LEP questionnaire would, on average, be at the same achievement level as students with disabilities and English language learners who participated in NAEP in that state and had the same demographics and the same SD/LEP questionnaire profile.

The method for computing full-population estimates is straightforward. Plausible values are estimated for each excluded student, and these values are used along with the plausible values of included students to obtain state-level statistics. Estimation of plausible values for achievement of excluded students consists of two steps.

- Finding the combination of information on the teacher assessment form and demographic information that best accounts for variation in achievement of included students with disabilities and English language learners. At the same time, estimate the amount of error in that prediction.
- Combining the information in the same manner for excluded students to obtain a mean estimate for each student profile. Generate (five) plausible values for each student profile by adding to the mean estimate a random variable with the appropriate level of variation.

This method can be used to generate full-population estimates, either using the score on accommodated tests or not. It would work equally well after setting accommodated scores to *missing*. Because NAEP is currently using accommodated scores, the full-population estimates presented here treat accommodated scores as valid indicators of achievement. The procedure was carried out separately for each grade and subject each year.

In 2004, the NAEP Quality Assurance contractor, the Human Resources Research Organization (HumRRO), tested the methodology used in this report to estimate the performance of the excluded students for sensitivity to violation of assumptions

(Wise et al., 2004). Overall, under the assumptions of the model, the full population estimates were unbiased. Violations of these assumptions led to slightly biased estimates which, at the jurisdiction level, were considered negligible.

The Education Testing Service (ETS) has recently developed an alternative approach to address the exclusion problem. ETS's approach is also an imputation procedure; it is based on the same basic assumptions used by AIR, with the only substantive difference being the inclusion of the school's state assessment score variable in the imputation model.[5] When both approaches were compared (Wise et al., 2006), their performances were equivalent. When model assumptions were violated, the ETS estimates were slightly less biased but, overall, the two approaches produced similar standard error estimates (Wise et al., 2006).[6]

The overall conclusion is that "the degree of bias in mean estimates generated from the FPE method was quite small and represented a significant improvement over simply ignoring excluded students even when excluded students' achievement levels were much lower than predicted by background information."

5. AIR deliberately excluded that variable in order to eliminate the argument that NAEP's FPE imputation might be based on something other than NAEP. For example, using state assessment results that include accommodations not allowed by NAEP may negatively impact the credibility of NAEP estimates for the excluded students.
6. The small differences between the two models seem to be mostly related to the inclusion of school's state assessment score variable in the ETS model.