NISS

# Projecting to the NAEP Scale: Results from the North Carolina End-of-Grade Testing Program

Valerie S. L. Williams, Kathleen Billeaud, Lori A. Davis, David Thissen, and Eleanor E. Sanford

Technical Report Number 34
June, 1995

National Institute of Statistical Sciences
19 T. W. Alexander Drive
PO Box 14006
Research Triangle Park, NC 27709-4006
www.niss.org

# Projecting to the *NAEP* Scale: Results from the North Carolina End-of-Grade Testing Program

Valerie S. L. Williams
*National Institute of Statistical Sciences*
Kathleen Billeaud
Lori A. Davis
David Thissen
*The University of North Carolina at Chapel Hill*
Eleanor E. Sanford
*North Carolina Department of Public Instruction*

ABSTRACT

Data from the North Carolina End-of-Grade test of eighth-grade mathematics are used to estimate the achievement results that would have been obtained from the National Assessment of Educational Progress (*NAEP*) 1994 Trial State Assessment, had it been administered. Linear regression models are used to develop projection equations to predict state *NAEP* results in the future. Standard errors of the parameter estimates are obtained using a bootstrap resampling technique. As an illustration, the projection equation is applied to data from North Carolina school districts to obtain estimates of district-level performance. As expected, there is substantial variability among the school districts, with a range of average (projected) scores comparable to that for the range of 1992 state *NAEP* averages.

---

Research supported in part through the North Carolina Department of Public Instruction, NSF DMS-9208758, and NCES/NSF through NSF RED-9350005. We are grateful to William Brown, Gary Phillips, and Eugene Johnson for their roles in generating the data reported here, and thank Bruce Bloxom, Jan de Leeuw, Bradley Efron, Paul Holland, Lyle V. Jones, John Mazzeo, Donald B. Rubin, John W. Tukey, and Paul Williams for helpful suggestions in the course of this research. Any errors that remain are, of course, our own.

## Projecting to the *NAEP* Scale: Results from the North Carolina End-of-Grade Testing Program

The *Goals 2000: Educate America Act* requires an instrument to assess the effects on student performance of education reform, and to monitor progress with respect to consensual national achievement standards. One approach to fulfilling this requirement is the establishment of linkages between state testing programs and a common metric of growth and change, such as the scale used for the National Assessment of Educational Progress (*NAEP*). With such linkages, results from more frequently-administered state assessments could be translated into estimates of results that would have been obtained had the *NAEP* Trial State Assessment (*TSA*) been administered. This would reduce the reliance on a national testing program such as the *TSA* for purposes of tracking student achievement, and is consistent with the recommendation of the National Council on Education Standards and Testing (1992), to facilitate the comparability of student outcomes across assessment instruments, across different education programs, and across states or other jurisdictions.

Not only could linkages serve to estimate state-*NAEP* results, but they could also provide comparable measures at the level of the local school district, or possibly at the level of the school building. While Congress recently has removed the prohibition against reporting *NAEP* results below the state level, neither the *NAEP* nor the *TSA* sampling designs currently support valid inferences for students, schools, or even school districts. Although Selden makes "the case for district- and school-level results from *NAEP*" (in Glaser & Linn, 1992), he concludes that:

> "If linking became available and economically feasible, it could be expected that states would use it to maintain particular features of, and purposes for, their testing programs, while tying into *NAEP* and generating *NAEP* scores for schools and districts. It would appear to be in the interest of assessment for states to be encouraged and enabled to develop statewide systems which are distinctive and creative, while tying into a national assessment system that provides local schools and

systems valuable data in a common national currency" (p. 96).

**An overview of test linkage**

Both Mislevy (1992) and Linn (1993) describe three levels of linkage that are potentially useful for re-expressing the results of locally-administered tests on the *NAEP* scale: *Equating*, *calibration*, and *projection* are the terms used to describe these three statistical approaches to linking tests. Test equating is the most rigorous form of linking. If two tests are equated, it is "a matter of indifference to applicants at every given ability level ... whether they are to take test $x$ or test $y$" (Lord, 1980, p. 195), and scores are interchangeable for any use of the tests. In practice, valid equating is usually limited to alternate forms of tests constructed from the same test specifications (Mislevy, 1992). Application of the most commonly used equating procedure, equipercentile equating, demands little in the way of data — only the score distributions for the two tests in the same population. It is therefore possible that other tests can be "equated" to the *NAEP* scale using published data; however, use of those data does not allow any direct evaluation of the accuracy or — perhaps, more importantly — the meaningfulness of the equating. While there have been attempts to equate other tests to the *NAEP* scale (Ercikan, 1993; Linn & Kiplinger, 1993), such equating does not provide fully satisfactory results. For example, with respect to the Linn and Kiplinger study, "in general, the equating did not hold, and analyses showed differences larger than chance along most parts of the achievement distribution" (Glaser & Linn, 1993, p. 127). Bond and Jaeger (1993) cite the lack of congruence in skill and content coverage as an important reason for the distortions found in many test equating studies.

Calibration is a somewhat less demanding statistical procedure for linking tests; it is usually based on the models of item response theory (IRT). Calibration is used to provide comparable scores on tests that "measure the same thing" (Mislevy, 1992, p. 22) but possibly with different degrees of precision, as may be the case with short and long forms of a test — revised forms of *NAEP* are calibrated with respect to each other, as are alternate forms of the North Carolina End-of-Grade tests. IRT calibration requires item response data from a sample of examinees administered both tests;

probably because item-level data from *NAEP* tests are made available only in special situations, there have been no attempts to calibrate other tests to *NAEP*.

If two tests are separately constructed to measure eighth-grade mathematics proficiency, there is very likely a strong positive relation between the scores. This will be true even if the content specifications and administration procedures are sufficiently different that the two tests do not assess exactly the same proficiency and cannot be reasonably equated, or even calibrated. Projection makes use of an empirical relation between scores on tests that do not measure exactly the same thing to predict the distribution of one test (e.g., *NAEP*) from the distribution of scores on another test (e.g., a state assessment) — Linn (1993) refers to this process as *prediction*. Although there may be differences in the correlation between scores across subpopulations, an empirically-estimated bivariate relation between the test scores, and the known marginal distribution of the scores on test $x$ (within subpopulations, if necessary), can nevertheless be used to infer the marginal distribution for test $y$. This procedure may not be useful as a method for constructing translation tables for individual scores because the result is not a single score or point estimate but a distribution of possible outcomes; in addition, the score translation may differ for subpopulations. However, the projection methodology can provide statistics useful for description and policymaking, e.g., an average or the proportion of students achieving at a given level.

While Mislevy (1992) and Linn (1993) provide thorough descriptions of the theoretical background necessary for various levels of linking, no examples are currently available for states to follow in linking locally-constructed tests to the *NAEP* scale. Kentucky is in the process of linking its state assessment to the *NAEP* achievement levels; however, Kentucky's test is scored categorically into four proficiency levels, so that model will not be widely used with local assessments that have scaled scores. Bloxom, Pashley, Nicewander, and Yan (1995) have described a linkage of scaled scores on the Armed Services Vocational Aptitude Battery with *NAEP*; however, both the data and the analytic procedures used in that effort are more complex than are common for state assessments, primarily because it involved a

number of subscale scores and multiple imputations of examinee proficiency estimates.

The present investigation reports the procedures and results of one successful attempt at linking a statewide assessment program to the *NAEP* scale using projection methodology. This study provides a practical model and explicates a set of procedures that can be followed in linking related but disparate tests.

## DEVELOPMENT OF THE NC-*NAEP* LINKAGE

The North Carolina Department of Public Instruction has developed a comprehensive academic testing program for grades 3 through 8, the End-of-Grade (EOG) tests. These tests assess the achievement of public school children in mathematics, reading, and social studies (a science test is under development). The score scale for mathematics, vertically equated to describe the performance of students in grades 3 through 8, ranges from about 100 to about 200, with an eighth-grade average of approximately 168 in 1993 and 169 in 1994; the within-grade standard deviation of the scaled scores is about 10. The *NAEP* mathematics scale ranges from 0 to 500, with an eighth-grade mean of approximately 262 for the nation in 1990 and 266 in 1992; the eighth-grade standard deviation is about 37 for the nation, and about 34 within North Carolina. In 1990, North Carolina eighth-graders averaged 250 on the *NAEP TSA* in mathematics, and in 1992, 258.

Mathematics proficiency, as measured by the *NAEP* exercises, is not identical to mathematics proficiency as measured by the EOG tests. The *NAEP* instrument is closely aligned with the *1991 Curriculum and Evaluation Standards for School Mathematics* established by the National Council of Teachers of Mathematics (NCTM); the EOG mathematics test was carefully designed to reflect the state curriculum objectives, the *Standard Course of Study*, which are based on the NCTM standards but delineated for each grade level. As illustrated in Table 1 by the percentages of items measuring the various objectives, there is considerable overlap in the content frameworks of the two tests.

### The sample

Eighth-grade examinees were selected in a two-stage sampling design

Table 1.

The percentages of items specified for the eighth-grade mathematics test and the number of items on the linking forms, within *NAEP* subscales and EOG objectives.

| ***NAEP*** | | | **EOG** | | |
|---|---|---|---|---|---|
| | Percentage specified | Number on linking test | | Percentage specified | Number on linking test |
| Numbers and Operations | 30 | 12 | Numeration | 11 | 5 |
| | | | Computation | 26 | 8 |
| | | | Problem-solving | 15 | 6 |
| Geometry | 20 | 5 | Geometry | 10 | 4 |
| Algebra and Functions | 20 | 4 | Pre-algebra | 15 | 7 |
| Measurement | 15 | 11 | Measurement | 10 | 4 |
| Data Analysis, Statistics, and Probability | 15 | 6 | Statistics | 12 | 6 |
| Total | 100 | 38 | | ~100 | 40 |

formulated by Westat, Inc., the sampling subcontractor for the *NAEP TSA*, for what would have been the 1994 administration of the *NAEP TSA* had it been funded by Congress. The primary sampling unit is the school: 103 schools were drawn, and 99 of these participated. A target sample of 30 students was randomly selected in each of the schools; actual counts ranged from 21 to 33 participants. Student exclusions were based on *NAEP* conventions of limited English proficiency status and individualized education plan allowances, resulting in within-school participation rates ranging from 70% to 100%. This compares favorably to previous *NAEP* participation rates reported for North Carolina students.

A total of 2824 students were tested. Table 2 contains the crosstabulated frequencies of students by sex and ethnic classification. The numbers in the "Native American," "Hispanic," and "Asian/Pacific Islander" ethnic classification categories were inadequate for separate projections. Two ethnic classifications reflecting relative educational advantagement were created for the projection analyses: BHN ("Black," "Hispanic," and "Native American" examinees) and WA ("White," "Asian/Pacific Islander," and "Other" examinees).

Table 2.

Sex and ethnic classification of linking sample.

| | Female | Male |
|---|---|---|
| Asian/Pacific Islander | 17 | 20 |
| Black | 462 | 431 |
| Hispanic | 16 | 18 |
| Native American | 18 | 11 |
| White | 906 | 889 |
| Other | 19 | 16 |
| Total | 1438 | 1385 |

*Note*: One black examinee did not report sex.

**Data collection**

The test administered in February 1994 contained 78 items, including a short

form of the North Carolina End-of-Grade mathematics test for grade 8 (40 multiple-choice items) and two blocks of released 1992 *NAEP* mathematics items (38 items: 29 multiple-choice and 9 free-response). As shown in Table 1, the EOG items represented each of the seven objectives about equally; similarly, the *NAEP* test items represented all five mathematics subscales. Coefficient alpha for the summed scores of the 38 *NAEP* items is $\alpha = 0.88$, and $\alpha = 0.82$ for the 40 EOG items. The reliability of the combined 78-item test is $\alpha = 0.91$. Approximately half of the test booklets (Form A; $N = 1330$) contained the *NAEP* test first, followed by the EOG test, and the rest of the booklets (Form B; $N = 1493$) contained the EOG test followed by the *NAEP* test.

Examination of the simple summed scores reveals differences between Forms A and B, and between population subgroups. The differences between *NAEP* and EOG summed score averages for the two orders of administration indicate that there is an order effect: The averages for EOG are 17.2 (Form A) and 16.6 (B), while the *NAEP* averages are 22.4 (Form A) and 21.2 (B); the standard deviations are about 7 points. When students took the EOG test first (Form B), they did worse on both tests. This is not entirely unexpected given that the EOG test is much more difficult and when administered first, probably led to fatigue or reduced motivation, producing lower scores. For this sample, the average proportion-correct is 0.42 for the EOG test compared to an average proportion-correct of 0.57 for the *NAEP* test.

As has been found in the past in the North Carolina testing program, females scored somewhat better on the EOG tests (the averaged summed score equals 17.2) than males (the averaged summed score is 16.5). However, on the *NAEP* test this difference was not observed; as shown in Table 3, females scored the same as males (21.8). The difference between the performance of black students and white students is greater on the *NAEP* test ($\bar{X}_{\text{Black}} = 17.2$, $\bar{X}_{\text{White}} = 24.1$) than on the EOG test ($\bar{X}_{\text{Black}} = 13.6$, $\bar{X}_{\text{White}} = 18.4$) — this result may be attributable to the greater difficulty of the EOG test.

Table 3.

Means and standard deviations of summed scores.

| | *NAEP* | EOG |
|---|---|---|
| Females | 21.8 (7.0) | 17.2 (6.5) |
| Males | 21.8 (7.1) | 16.5 (6.7) |
| Black | 17.2 (6.0) | 13.6 (4.9) |
| White | 24.1 (6.4) | 18.4 (6.7) |
| Total | 21.80 (7.06) | 16.85 (6.59) |

The procedures used to project the *NAEP* scaled score distribution require the IRT item parameters for the two blocks of items that were administered to the linkage sample. Published item parameter estimates from the *NAEP TSA* documentation are based on separate within-subscale item calibrations using unidimensional two- and three-parameter logistic item response models. For this study, the Educational Testing Service (ETS) provided estimates of *a*, *b*, and *c* parameters for each item from a unidimensional three-parameter logistic model with proficiency defined as the principal axis obtained in an item analysis of the entire 1990 and 1992 *NAEP* item pool.

**Analyses**

The NC-*NAEP* linkage analyses proceeded through two phases:

- Selection of a model for *NAEP* averages and standard deviations, conditional on EOG scores and background variables.
- Bootstrap computation of standard errors for the regression coefficients.

This report describes the results from each phase of the NC-*NAEP* linkage analysis, as well as the decisions made at each step.

**Selection of a model for *NAEP* averages and standard deviations.** For each student, a *NAEP* posterior distribution is obtained based on the individual response pattern, the population distribution, and the IRT parameter estimates provided by ETS. The prior, also provided by ETS, is a non-Gaussian histogram for the 1992 national *NAEP*, derived from an analysis of the 1990 and 1992 *NAEP* tests. Each examinee's posterior distribution is represented by a probability polygon defined by the relative

likelihood of that examinee's response pattern at 37 equally-spaced values of proficiency, the *quadrature points*. These distributions were rescaled, so that the height at each quadrature point is a proportion of 1.0 and each individual's posterior sums to 1.0. The sum of these distributions, weighted by the sampling weights, is the sample estimate of the 1994 distribution.

The EOG summed scores are transformed to EOG scaled scores. Students are categorized into groups based on sex, ethnic classification, and EOG scaled score. By summing the weighted posteriors for each sex × ethnic classification × EOG score combination, four distributions of *NAEP* scores for each EOG scaled score category are created. Early analyses indicated that the conditional distributions did not differ reliably by sex; in what follows, we consider only the subdivision by ethnic classification. Figure 1 shows the conditional posterior distributions for three EOG scaled scores.

The projection equations fit the posterior mean of each ethnic classification × EOG score category as the dependent variable; this is the mean of the posterior distribution created by summing all the individual posteriors for each examinee in an ethnic classification × EOG score category. The predictors are ethnic classification (dummy-coded BHN = 0 and WA = 1) and EOG score category, centered by subtracting 165, the mean EOG scaled score for the linkage sample. The standard deviations of the ethnic classification × EOG score category posteriors are predicted from EOG score category only. Weighted least squares regression analysis, in which the ethnic classification × EOG score category subgroupings are weighted by the number of students in each subgrouping, produced the parameter estimates shown in Table 4. Inclusion of the ethnic classification × EOG score category interaction did not contribute significantly to the prediction of the means of the posteriors. The means for all the score categories, and the two regression lines, are shown in Figure 2; the standard deviations are similarly shown.

**Bootstrap computation of standard errors.** The nested sampling design precludes inferences based on estimates of uncertainty calculated according to assumptions of simple random sampling. Standard errors for the regression

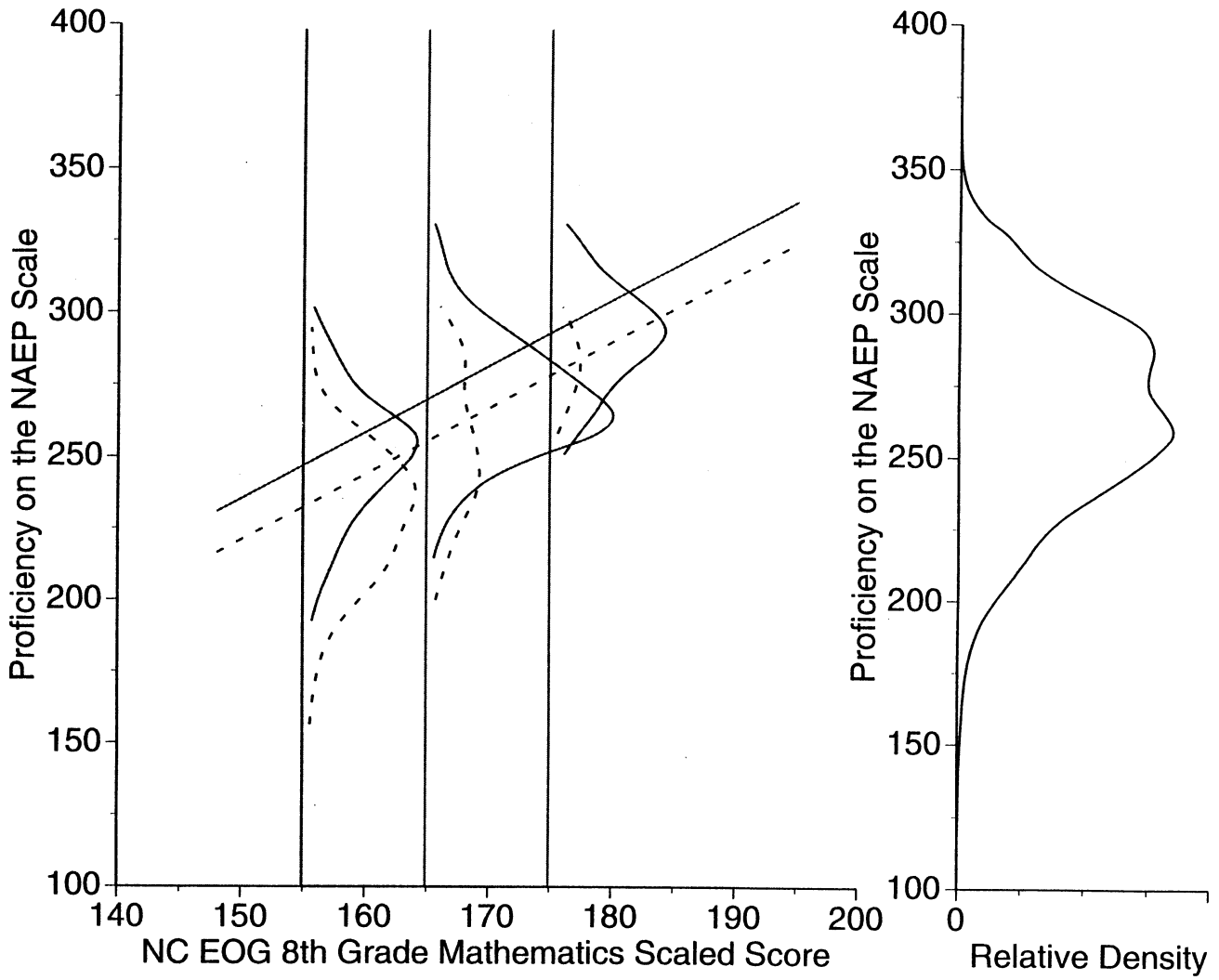
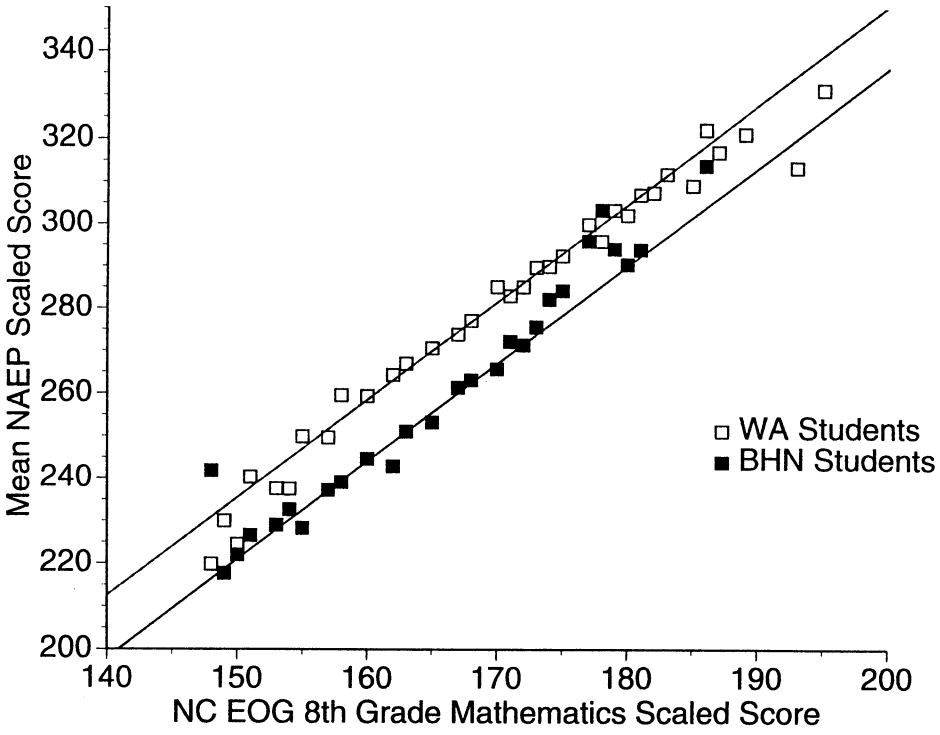



Figure 1. Conditional posterior distributions of *NAEP* proficiency for three EOG scaled scores (155, 165, and 175), for examinees in the BHN (dashed lines) and WA (solid lines) ethnic classifications. Shown at the right is the total *NAEP* posterior distribution, the weighted sum of the conditional distributions for all scores (140 to 204).



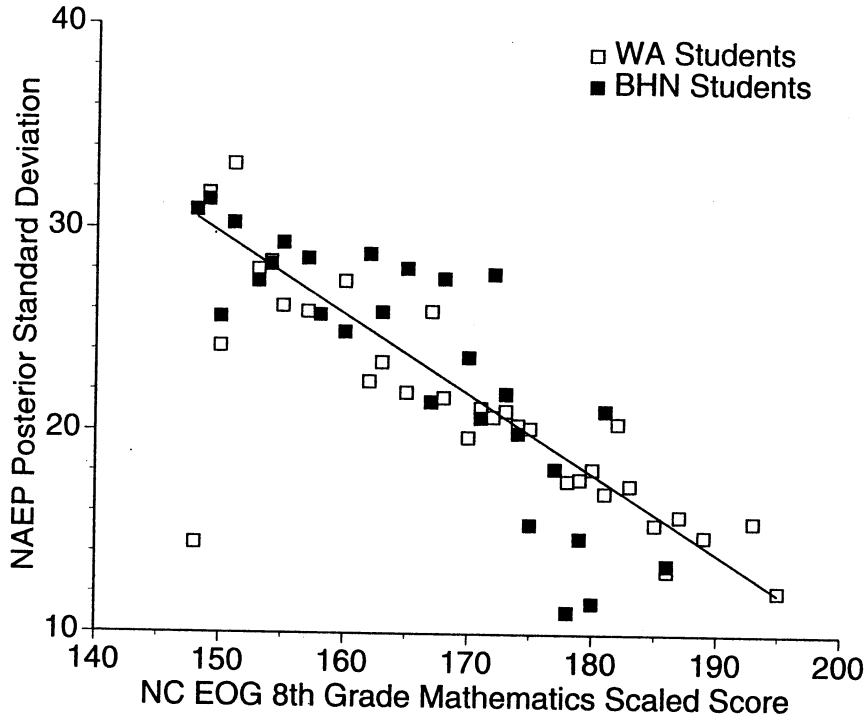




Figure 2. Means for each ethnic classification × EOG score category, and the fitted regression lines (above), and standard deviations for each ethnic classification × EOG score category, and the fitted regression line (below).

Table 4.

Parameter estimates from the weighted least squares regression model (for projecting February *NAEP* results from February EOG results).

For the prediction of ethnic classification × EOG score category posterior means:

| Variable | Coefficient | Standard error |
|---|---|---|
| Centercept | 255.80 | 0.91 |
| WA | 14.11 | 1.10 |
| EOGscore − 165 | 2.29 | 0.06 |

For the ethnic classification × EOG score category posterior standard deviations:

| Variable | Coefficient | Standard error |
|---|---|---|
| Centercept | 23.60 | 0.32 |
| EOGscore − 165 | −0.39 | 0.03 |

coefficients were computed using a bootstrap procedure described by Sitter (1992a, 1992b). The bootstrap plan included finite population corrections at the first and second sampling stages, for school and for student-within-school. In practice, the finite population correction resamples $n^*$ schools selected with replacement from the 99 schools, and $m^*$ students selected with replacement from each school. According to Sitter (1992a):

$$n^* = (n-1)/(1-f_1)$$

where $n$ is the number of clusters in the sample, $N$ is the total number of clusters in the population, and $f_1 = n/N$, and

$$m_i^* = (m_i-1)N/(1-f_{2i})n^*$$

where $m_i$ is the number of students in the $i$th sample cluster, $M_i$ is the total number of students in the $i$th cluster, and $f_{2i} = m_i/M_i$ . There are $N = 658$ schools with eighth grades in North Carolina, and $n = 99$ schools are represented in the NC-*NAEP* linkage sample, resulting in $n^* = 115$ schools to be resampled. The sizes of the eighth-grade classes, $M_i$, range from 29 to 496, and the size of the school-level samples, $m_i$, range from 21 to 33; the adjusted school-level sample sizes for the bootstrap, $m_i^*$, range

from 24 to 3204, although the maximum was set to 300 to reduce computation.[1]

The bootstrap is operationalized in four steps as follows:

Step 1 From the set of 99 schools, 115 schools are randomly selected with replacement; by chance, some schools are unrepresented, some duplicated, some triplicated, etc.

Step 2 From each school, $m_i$* (between 24 and 300) students are randomly selected with replacement; again, some students are unrepresented, duplicated, etc.

Step 3 Using the data obtained in Step 2, the mean and standard deviation for each ethnic classification × EOG score category are calculated, and the projection equations computed to obtain the five regression coefficients.

Step 4 Steps 1 through 3 are repeated a total of 200 times, producing 200 estimates for each statistic.

To obtain the bootstrap estimate of a parameter, the mean of each set of 200 estimates is calculated, and the standard error of each of the statistics is the standard deviation computed from the sampling distribution. The bootstrap parameter estimates are not used, but the standard errors from the bootstrapped regression with finite population corrections appear in Table 4 with the weighted least squares regression coefficients.

Figure 3 shows the smoothed (Gaussian) posterior distributions of *NAEP* proficiency for three EOG summed scores, for BHN and WA examinees. The posteriors were approximated using a Gaussian distribution, with the mean obtained from the regression for the means, and the standard deviation obtained from the regression for the standard deviations.

**Projection of February *NAEP* results from the May EOG administration.** A second analysis projected the February *NAEP* results from the regular May administration of the EOG test. A total of 2313 students from the NC-*NAEP* linkage sample were matched with their May EOG scores; the average EOG increased about five points for this sample (to $\bar{X}$ = 169). For future prediction of *NAEP* from the regular administration of the EOG, parameter estimates were again obtained using

---

[1] For four of the 99 schools, $m_i$* exceeded 300; the actual values are 315, 409, 996, and 3204.

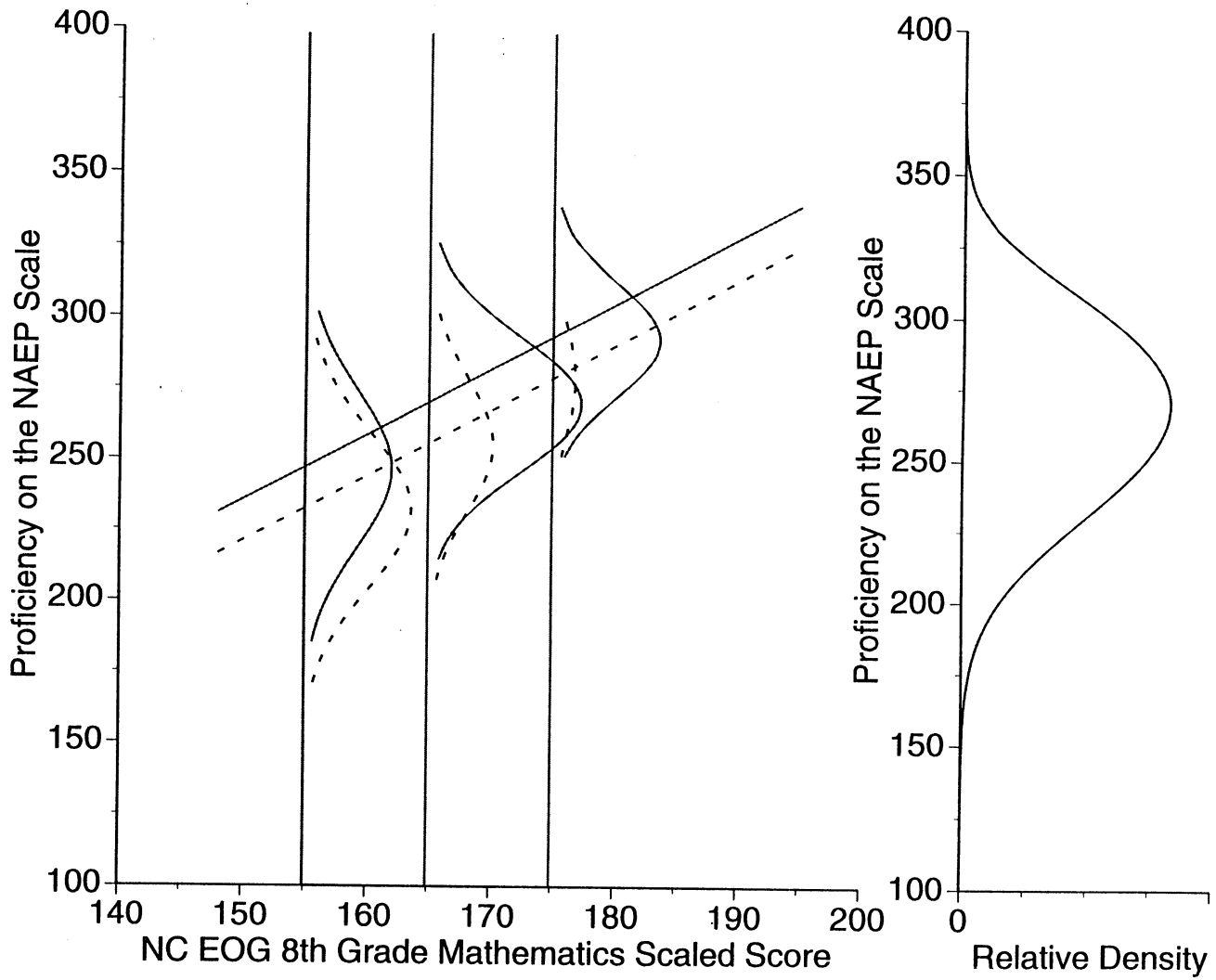



Figure 3. Smoothed (Gaussian) posterior distributions of *NAEP* proficiency for three EOG scaled scores (155, 165, and 175), for examinees in the BHN (dashed lines) and WA (solid lines) ethnic classifications. Shown at right is the total *NAEP* posterior distribution, the weighted sum of the conditional distributions for all scores (140 to 204).

weighted least squares, and standard errors for the parameter estimates were computed using the bootstrap procedure, as described above. Table 5 contains these regression coefficients and the standard errors for predicting February's *NAEP* results from the May EOG test administration; the values differ very little from those in Table 4.

Table 5.

Parameter estimates from the weighted least squares regression model for projecting February *NAEP* results from May EOG results.

For the prediction of ethnic classification × EOG score category posterior means:

| Variable | Coefficient | Standard error |
|---|---|---|
| Centercept | 259.96 | 0.87 |
| WA | 9.40 | 1.08 |
| EOGscore – 169 | 2.25 | 0.04 |

For the ethnic classification × EOG score category posterior standard deviations:

| Variable | Coefficient | Standard error |
|---|---|---|
| Centercept | 21.12 | 0.33 |
| EOGscore – 169 | –0.30 | 0.03 |

**Computation of standard errors for the statistics derived from the projection.** The empirical bootstrap procedure was used to compute the complete covariance matrices of the five regression parameters involved in the projections; the standard errors of the regression coefficients reported in Tables 4 and 5 are the square roots of the diagonal elements of those matrices. Simulation is used to compute estimates of the standard errors of the statistics derived from the projection, such as the projected percentiles. We simulate the effects that the use of different samples to develop the projection might have on the statistics derived from the projection.

To accomplish the simulation, it is assumed that the five regression coefficients are drawn from a multivariate normal distribution, with the mean equal to the estimates and the covariance matrix computed with the empirical bootstrap. Then 200 projections are done, using as the five regression parameters random draws from that

multivariate normal distribution. The standard deviations of the derived statistics, such as the percentiles, computed over the 200 simulated projections, are reported as the standard errors of the derived statistics. Efron and Tibshirani (1993, p. 53) refer to such simulation for computing standard errors as the "parametric bootstrap," and describe the motivation for the procedure as a means "to provide answers in problems where no textbook formulae exist" (p. 55).

**1994 state results**

The 1994 *NAEP TSA* results for North Carolina were obtained directly from the linkage sample. Subsequently, when the data from the statewide census administration of the EOG test became available, the projection equations summarized in Table 5 were developed, and the data from all 82,657 eighth-grade students were used to project (or, in this case, postdict) the February *NAEP* results.

Figure 4 shows the estimated proficiency distribution from the projection compared to that obtained directly from the *NAEP* administration. The distributions correspond closely, with the visually salient exception that the *NAEP* distribution, computed directly as the sum of the individual IRT posterior distributions, is slightly bimodal, while the projected approximation is unimodal. Experience with these data indicate that the (apparent) bimodality for the proficiency distribution disappears if a Gaussian approximation for the IRT posteriors is incorporated at any step in the computation. In the projection, that approximation is introduced for the inferred *NAEP* proficiency distributions for the ethnic classification $\times$ EOG score category groups. In the usual analysis of *NAEP TSA* data, the Gaussian approximation is used when plausible values are generated; *NAEP TSA* results based on the IRT posterior distributions themselves are not generally reported, so the bimodality is usually not apparent.

*NAEP TSA* results are most commonly reported as a set of quantiles of the distribution. Table 6 shows the values of the percentiles typically reported, as observed in the 1994 special administration of NC-*NAEP* to the linkage sample of 2824 students, and as projected from the (near) population of 82,657. Six of the seven percentiles from the projection are within one standard error of the original sample

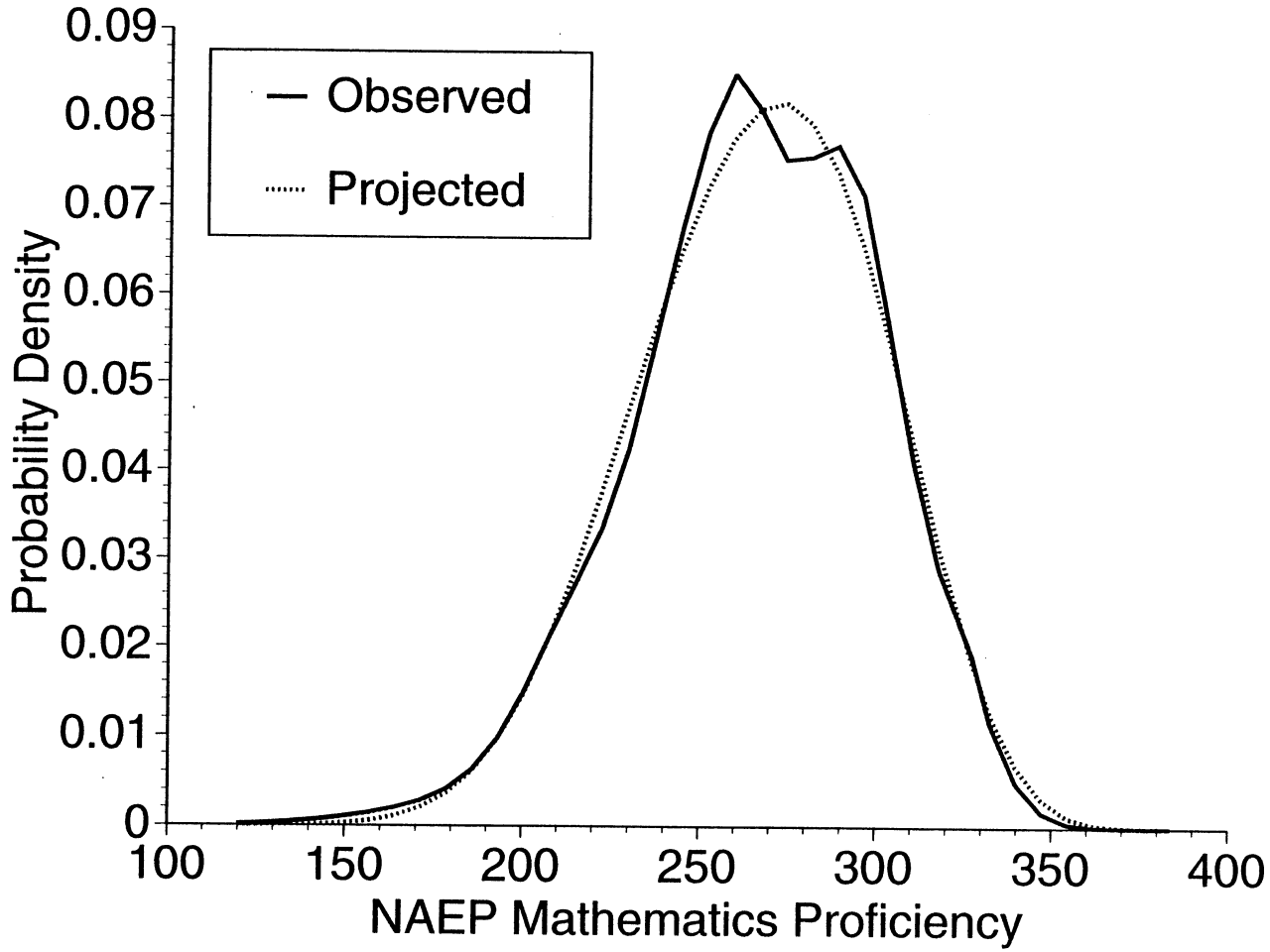



Figure 4. The distribution of North Carolina eighth-grade mathematics proficiency from the NC-*NAEP* administration (solid line) and the projection from the statewide administration of the EOG test (dashed line).

values, and the seventh is well within two standard errors. It should be noted that the standard errors for the projected values are smaller than those computed with the original sample. These standard errors take into account only the sampling variation in the projection itself: Because the data from which the projection is done are population values, there is no sampling variation from that source. Measuring the population with the wrong test results in less sampling variation than making an inference to the population with data from the right test but using a smaller sample. (There are, of course, both systematic and random sources of error that are not captured in sampling variation; those sources of error are not reflected in either set of standard errors.)

Table 6.
Observed and projected percentiles for the distribution of mathematics proficiency for the North Carolina eighth-grade students (bootstrap standard errors are shown in parentheses).

| | 5th | 10th | 25th | 50th | 75th | 90th | 95th |
|---|---|---|---|---|---|---|---|
| Observed | 206 (2.0) | 220 (2.0) | 244 (1.7) | 267 (1.7) | 291 (1.3) | 308 (1.4) | 319 (1.3) |
| Predicted | 208 (1.4) | 221 (1.1) | 243 (0.8) | 268 (0.6) | 291 (0.5) | 310 (0.6) | 320 (0.6) |

When the data from the 1995 administration of the EOG eighth-grade mathematics test become available, it will be possible to project the state's 1995 *NAEP TSA* results. Moreover, with the next administration of an eighth-grade mathematics *TSA*, the results of the present NC-*NAEP* projection will be evaluated for accuracy.

**Projection of results for school districts**

As an illustration, the average *NAEP* scores for North Carolina's 119 school districts were projected. For each school district, students are grouped into ethnic classification × EOG scaled score categories, and the *NAEP* posterior mean of the

ethnic classification × EOG score category is predicted using the linear model:

$$\bar{X}_{NAEP} = 259.96 + 9.40\text{WA} + 2.25(\text{EOGscore} - 169) .$$

The standard deviations of each ethnic classification × EOG score category posterior are predicted from EOG score category:

$$s_{NAEP} = 21.12 - 0.30(\text{EOGscore} - 169) .$$

The estimates of district-level mathematics performance show a large amount of variability within the state of North Carolina, with school district averages ranging from a minimum of 239 to 286. This maximum value represents mathematics performance comparable to the highest state *NAEP* averages. For example, in 1992, the eighth-graders in Iowa and North Dakota averaged 283. The lower value indicates poor student performance similar to that in states such as Mississippi ($\bar{X}$ = 246) and Louisiana ($\bar{X}$ = 249), or in the District of Columbia and Guam ($\bar{X}$ = 234).

## Discussion

Projection methodology, as Mislevy (1992) says, is "rather precarious" (p. 63), largely because it relies on the empirical relation between qualitatively different evidence about the proficiencies of individuals and groups. One source of statistical uncertainty is model misspecification: Either the IRT or the (linear) regression models could be incorrect, or the assumed population distribution could be erroneous. There is uncertainty due to the sampling error associated with the calibration sample, as well as the error in the projection sample. (However, in the latter case, the error associated with the projection sample may be negligible — for the North Carolina testing program, the entire population is tested.)

Each *NAEP* mathematics item is associated with one of five subscales, with item parameter estimates based on within-subscale calibrations. For this projection, however, a unidimensional item response model was used with the *NAEP* items. This captures the main pattern of response variation and results in a summarization of the distribution of overall mathematics proficiency (Mislevy, Johnson, & Muraki, 1992), as opposed to separate — but highly correlated — subscale estimates.

It should be noted that the items in a regular *NAEP TSA* are administered in a

balanced incomplete block design where different blocks of *secure* items appear on different test forms which are spiralled within classroom. The NC-*NAEP* linkage administration used a fixed (counter-balanced) test containing *released NAEP* 1992 items. The consequences of using released, instead of secure, *NAEP* items in the current testing environment are unknown.

ETS has developed the *plausible values* technology, or multiple imputations (Rubin, 1987), especially for analysis of *NAEP* data. Because too few items are administered to each examinee to accurately compute point estimates of examinee proficiencies for each mathematics subscore and composite score, a posterior distribution is used to represent each examinee's proficiency for each mathematics subscale and overall composite. For analytical purposes, each examinee is then assigned five plausible values for each subscale score; these are values randomly drawn from each examinee's posterior proficiency distribution. In contrast, the NC-*NAEP* linkage analyses used pointwise representations of the examinee's posterior distributions. In theory, this computational difference should have no effect on the results.

The population distribution used in the *NAEP TSA* is conditioned on a large number of principal components of variables collected from an extensive background questionnaire administered with the *NAEP* cognitive tests. The choice of conditioning variables affects the size of the root mean squared error of all parameter estimates in a predictable manner, i.e., greater population variance will translate into larger error variance. These conditioning variables are like additional items on the test. In the NC-*NAEP* linkage, no conditioning background variables were used.

The standard errors reported for *NAEP* tests are jackknifed estimates of uncertainty; the NC-*NAEP* linkage produced standard errors for the regression coefficients by computing the standard deviations of the bootstrapped distributions. Longford (1995) found that jackknifed estimates are biased, possibly because they neglect the within-cluster variability. The bootstrap technique used here includes both within-cluster variability and finite sample corrections, but raises questions that are beyond the scope of this study: The *NAEP* jackknife standard errors include a

variance component attributed to the fact that proficiency is an unobserved latent variable (Mislevy, Johnson, & Muraki, 1992, p. 146), while the bootstrap/simulation procedures used in the present study do not — what is the place of this source of variance in projection? Are the sampling distributions of the statistics derived from the projection adequately described as Gaussian distributions, with variance equal to the squared simulated standard errors? Is the finite-population-corrected bootstrap (Sitter, 1992a, 1992b) an optimal way to evaluate sampling variation for complex samples of the kind used for *NAEP*, or might a model-based approach (e.g., Longford, 1995) be more accurate and useful? Answers to these questions require a more precise specification of the sources of uncertainty that are to be included in the description of variation for statistics derived from *NAEP* and other such assessments, as well as further research.

Eighth-grade students are, of course, encouraged to perform as best they can; however, the 1990 and 1992 administrations of the *NAEP TSA* were likely perceived by sampled students as being of little importance because they were aware that individual test scores would not be reported. In this study, the EOG test also can be characterized as relatively "low-stakes" in that students are informed that the scores they receive on the tests do not directly affect their class grades; however, students *do* receive score reports that describe individual performance. Motivational differences are cited by both Bloxom et al. (1995) and Ercikan (1993) as possible reasons for the failure of linkages, and the importance of motivational factors should not be underestimated here.

Another practical issue involves the time frame for which the empirical relation between the *NAEP* and EOG can be expected to hold, and how frequently the projection equations must be recomputed. In light of evidence indicating that school curriculum can be driven substantially by high-stakes tests (cf. Koretz, Linn, Dunbar, & Shepard, 1991), it is unreasonable to expect that the current linkage will remain stable for an indefinite period of time, especially if either the *NAEP* or EOG tests come to play a role in school and school district accountability.

Each of the above concerns should be considered challenges to the

interpretation of the NC-*NAEP* projection results. They are important issues that remain to be resolved by further study.

**Conclusions**

Because of the great expense involved in expanding *NAEP* to provide scores below the state level, a network of state-*NAEP* linkages may provide a more feasible solution for *NAEP* score reporting at the school district level. North Carolina has developed a student achievement testing program which also serves as one mechanism for school district accountability. The NC-*NAEP* linkage will not only permit the state to make district-level comparisons to national data, but it also allows comparisons of school district progress with respect to national standards. *NAEP* linkages would also facilitate state- and district-level comparisons with international results.

References

Bond, L., & Jaeger, R. M. (1993). Final report on the judged congruence between various statewide assessment tests in mathematics and the 1990 National Assessment of Educational Progress content frame for grade 8 mathematics. Greensboro, NC: Center for Educational Research and Evaluation, University of North Carolina at Greensboro.

Bloxom, B., Pashley, P., Nicewander, A., & Yan, D. (1995). Linking to a large-scale assessment: An empirical evaluation. *Journal of Educational and Behavioral Statistics, 20,* 1-26.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap.* New York: Chapman & Hall.

Ercikan, K. (1993). Predicting *NAEP.* Unpublished manuscript. Monterey, CA: CTB Macmillan/McGraw-Hill.

Glaser, R., & Linn, R. (Eds.). (1992). *Assessing student achievement in the states.* Stanford, CA: National Academy of Education.

Glaser, R., & Linn, R. (Eds.). (1993). *The Trial State Assessment: Prospects and realities.* Stanford, CA: National Academy of Education.

Koretz, D., Linn, R., Dunbar, S., & Shepard, L. (1991). The effects of high-stakes testing on achievement: Preliminary findings about generalization across tests. Paper presented at the Annual Meeting of the American Educational Research Association and the National Council on Measurement in Education, Chicago, IL.

Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education, 6,* 83-102.

Linn, R. L., & Kiplinger, V. L. (1993). *Linking statewide tests to the National Assessment of Educational Progress: Stability of results.* Boulder, CO: Center for Research on Evaluation, Standards, and Student Testing.

Longford, N. L. (1995). *Model-based methods for analysis of data from the 1990 NAEP Trial State Assessment.* Research and Development Report. Washington, DC: U.S. Department of Education, Office of Educational Research and

Improvement, National Center for Education Statistics.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects.* Princeton, NJ: Educational Testing Service.

Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Scaling procedures in *NAEP. Journal of Educational Statistics, 17,* 131-154.

National Council on Education Standards and Testing. (1992). *Raising standards for American education.* Washington, DC: Author.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* NY: Wiley.

Sitter, R. R. (1992a). A resampling procedure for complex survey data. *Journal of the American Statistical Association, 87,* 755-765.

Sitter, R. R. (1992b). Comparing three bootstrap methods for survey data. *The Canadian Journal of Statistics, 20,* 135-154.