

## Adjusting for Differences in Tests

Robert L. Linn

University of Colorado at Boulder

National Center for Research on Evaluation, Standards, and Student Testing

Paper prepared for a Symposium on the Use of School-Level Data for Evaluating Federal Education Programs,. Washington, DC: The Board on Testing and Assessment, The National Academies, December 9, 2005.

*Copyright © 2005 National Academy of Sciences. All rights reserved. No part of these pages, either text or image may be used for any purpose other than personal use. Reproduction, modification, storage in a retrieval system or retransmission, in any form or by any means, electronic, mechanical or otherwise, for reasons other than personal use, is strictly prohibited without prior written permission.*

*Opinions and statements included in the paper are solely those of the individual author(s), and are not necessarily adopted or endorsed or verified as accurate by the Planning Committee for the Symposium on the Use of School-Level Data in Evaluating Federal Education Programs or the National Academy of Sciences, including the National Academy of Engineering, Institute of Medicine, or National Research Council.*

## **Adjusting for Differences in Tests**

The desire to treat scores obtained from different tests as if they were interchangeable, or at least comparable in some more limited sense, is hardly new. While I was busily working to complete my dissertation, a symposium entitled “Equating Non-Parallel Test Scores” was held at the 1964 Annual meeting of the National Council on Measurement in Education. Participants in that symposium, Bill Angoff, John Flanagan, Roger Lennon, and E. F. Lindquist, clearly would be included in anyone’s Who’s Who of educational measurement. The four papers from that symposium became the lead articles in the first issue of the *Journal of Educational Measurement* (JEM).

All four authors (Angoff, 1964; Flanagan, 1964; Lennon, 1964; and Lindquist, 1964) expressed reservations about attempts to meet the demands for obtaining equated scores from non-parallel tests or efforts to produce conversion tables. They cautioned that it cannot be assumed that converted scores will behave just like those of the test whose scale is adopted. They expressed concerns that results are apt to be misinterpreted despite cautions to users about the limited senses in which scores may be considered comparable. They stressed that the conversions are apt to be specific to the groups of examinees used to develop the conversion function. It would appear, however, that the reservations of these leaders in the field of educational measurement did little, if anything, to stem the demand for conversions that would enable users to compare results of different tests on a common metric. Concordance tables linking the ACT and SAT, which Lindquist worried would lead to misinterpretations, have become commonplace, for example, and demands for comparable scores have continued to increase over the last four decades.

## ***K-12 Achievement Tests and Federal Evaluation Requirements***

Legislation enacted the year after the publication of the papers on equating non-parallel tests in JEM led to an increased appetite for treating the scores on achievement tests of different publishers as if they were interchangeable. The Elementary and Secondary Education Act (ESEA) of 1965, included requirements for the use of test results for the evaluation of programs supported by the Act. Title I of ESEA provided financial support for compensatory education to schools serving poor children. Testing requirements for Title I students were instituted as the result of congressional demands for evaluation and accountability. Initially the test requirements encouraged schools to administer standardized, norm-referenced achievement tests to Title I students and results were generally reported in terms of grade-equivalent scores. It soon became evident that grade-equivalent scales varied greatly from one test publisher to another and from one content area to another, making comparisons across school districts or across states impossible (for discussions of properties of grade-equivalent scores, see, for example, Angoff, 1971; Linn & Slinde, 1977; Petersen, Kolen, & Hoover, 1989).

Concerns about the lack of comparability of scores obtained from different norm-referenced tests stimulated the exploration of the possibility of equating tests of different publishers. A review of content on the various norm-referenced tests led to the conclusion that it was not sensible to try to equate mathematics tests, but that it did seem reasonable to launch a major data collection effort to equate the major reading tests used for students in grades 4, 5, and 6. Consequently, a major study, commonly referred to as the Anchor Test Study (ATS), was undertaken (Bianchini & Loret, 1974). The ATS had two

goals: (1) to equate the 7 most widely used norm-referenced reading tests<sup>1</sup> and (2) to obtain nationally representative norms for the tests.

By many measures the ATS was a successful undertaking. The study results provided a means of translating student scores obtained on any one of the reading tests at grades 4, 5, or 6 to the scale of any other test included in the study. Comparisons of equating functions for subgroups of students indicated that the equating functions were reasonably, albeit not perfectly, invariant across subpopulations of students. The re-standardization also created national norms that were of higher quality than those obtained by any single publisher. On the other hand, within a few years of the publication of the ATS results, all the test publishers had released new versions of their tests which made the ATS results far less useful. The new version of a test would first have to be converted to the scale of the old version using publisher results and then converted to the scale of one of the other tests using ATS results.

The ATS did lead to increased use of percentile ranks instead of grade-equivalent scores. Although percentile ranks avoided some of the problems caused by lack of comparability of grade-equivalent scores, percentile ranks have properties that are not well suited to statistical analyses. In an effort to deal with the limitations of percentile ranks, the U.S. Department of Education introduced the Title I Evaluation and Reporting System (TIERS) (Tallmadge & Wood, 1981). TIERS encouraged the administration of norm-referenced standardized tests to Title I students in both the fall and the spring. Programs were evaluated in terms of normal-curve equivalent scores (NCEs)<sup>2</sup>, which are normalized conversions of the publisher's percentile ranks obtained from their

---

<sup>1</sup> An eighth test was added to the initial list of 7 tests.

<sup>2</sup> NCEs are a normalized standard score translation of percentile ranks scaled in such a way that NCEs of 1, 50, and 99 are equal to percentile ranks of 1, 50, and 99, respectively.

norms. When NCEs from different tests were aggregated at the state or national level, it was simply assumed that the scores from different tests could be treated interchangeably. In other words, the possibility that the NCE scores of different publishers were not comparable was simply ignored when the NCE scores were aggregated. The aggregation ignoring possible differences is similar to using standard scores in a meta analysis, a practice that will be briefly considered later.

The assumption that NCE scores from different tests could be treated interchangeably, while supported to some extent at one point in time for reading tests by ATS results, was questionable for newer versions of reading tests which had to rely on publisher specific norms and even more questionable for mathematics tests. The assumption was rarely challenged, however, possibly because little use was made of the aggregate test results obtained from TIERS.

### ***Standards-Based Assessments***

Although there certainly were challenges to achieve comparability across the norm-referenced tests of different publishers in the 1970s and 1980s they pale in comparison to the challenges of the 1990s and those faced today. Several states (e.g., Kentucky and Maryland) adopted content standards and started developing standards-based assessments in the early 1990s. Some of these assessments made heavy use of constructed-response items rather than the multiple-choice test items found on traditional norm-referenced tests. Content standards, student performance standards, and standards-based assessments were central to the Clinton administration's education initiative explicated in the Goals 2000: Educate America Act. Use of a standards-based approach to assessment was reinforced by the requirements for Title I

evaluations mandated in the 1994 re-authorization of ESSA by the Improving America's Schools Act of 1994.

Those wanting to be able to compare scores on tests used by different states were soon faced with a relatively unique assessment in each state, rather than the 6 to 8, most commonly used norm-referenced achievement tests. The standards-based assessments were also more variable in content coverage and in format than the norm-referenced tests. Uses of test results and the stringency of performance standards varied from state-to-state. The apparent lack of comparability provided some of the motivation for President Clinton's proposal in 1997 to create a voluntary national test (VNT). The proposed VNT raised concerns among proponents of local control of curriculum that a national test, even if voluntary, would be tantamount to imposing a national curriculum.

There was a strong negative reaction to the VNT among some members of Congress. The negative reaction was led by Representative William Goodling who was then chair of the House Education and Workforce Committee. Representative Goodling wanted decisions about assessments to be left to the states. He addressed the desire for comparability by asking whether it might not be possible to let states use tests of their own choosing, but to somehow convert the scores on the different tests to a common scale (See Feuer, 2005, in press, chapter 2, for a more detailed discussion). Consequently, the National Research Council (NRC) was asked to investigate the possibility of converting scores on the myriad assessments used by different states to a common scale.

The NRC formed a study committee, chaired by Paul Holland, to address the question raised by Representative Goodling. The NRC committee studied the possibility of creating a single scale that could be used for reporting results for the various state tests and concluded that the answer to the question presented to the committee was simply “no”. Specifically the committee concluded that it is not feasible to compare “the full array of currently administered commercial and state achievement tests to one another, through the development of a single equivalency or linking scale” (Feuer, Holland, Green, Bertanthal, & Hemphill, 1999, p. 91). This conclusion was based on analyses that showed the assessments of different states varied so much in content, item format, conditions of administration, and consequences attached to the results that the linked scores could not be considered sufficiently comparable to justify the development of single equivalency scale.

A number of criteria must be met for an equating to yield strictly equivalent scores (see, for example, Angoff, 1971; Dorans & Holland, 2000; Holland, 2005; Linn, 1993; Mislevy, 1992). Dorans and Holland (2000), for example, identified the following five requirements for equating that enjoy a broad professional consensus:

- (a) *The Equal Construct Requirement.* Tests that measure different constructs should not be equated.
- (b) *The Equal Reliability Requirement.* Tests that measure the same construct but which differ in reliability should not be equated.
- (c) *The Symmetry Requirement.* The equating function for equating the scores of Y to those of X should be the inverse of the equating function for equating the scores of X to those of Y.

(d) *The Equity Requirement*: It ought to be a matter of indifference for an examinee to be tested by either one of the two tests that have been equated.

(e) *Population Invariance Requirement*: The choice of (sub) population used to compute the equating function between the scores on tests X and Y should not matter. In other words, the equating function used to link the scores of X and Y should be population invariant (pp. 282-283).

During the 1990s there were a number of attempts to link state assessments to NAEP (e.g., Ercikan, 1997; Linn & Kiplinger, 1995; McLaughlin, 1998; Waltman, 1997; Williams, Rosa, McLeod, Thissen, & Sanford, 1998). Thissen (2005) has recently provided a comprehensive review of those efforts. Based on his review, Thissen (2005) was led to the following conclusion. "While it is possible to construct linkages of statewide (or other) assessments to the NAEP scale, the linkages are (uniformly) not invariant over subpopulations [and] not particularly stable over time" (pp. 25-26 of typescript). Thus, there was a reasonable amount of evidence that equating requirement (e), invariance of results was not likely to be satisfied.

Careful consideration suggests that requirements (a), (b), and (d) are not satisfied either for assessments developed for use in different states. The content standards to which the assessments are supposed to be aligned may be similar for different states, but they are hardly identical. They have different emphases and as a consequence the specifications for the assessments are different making it hard to argue that the assessments of different states measure the same construct. The assessments vary in length and in the mix of different types of items. As a consequence, the assessments used by different

states vary in reliability. Furthermore, students whose teachers stress the state content standards and material covered by the state's assessment, including not only content but also item formats, clearly would not be indifferent to which assessment they took. Thus, in a strict sense, it seems clear that the NRC committee reached a justifiable conclusion.

Nonetheless, the committee chair, Paul Holland and some other members of the committee, e.g. Frederick Mosteller, expressed some dissatisfaction with the conclusion, in part, because of their natural predilection as statisticians to find solutions to practical problems and, in part, because of the lack of any quantitative basis for judging the degree to which the validity of interpretations of scores would be compromised by creating the linkages among the various assessments (see, Feuer, in press, chapter 2; Holland, 2005).

As is evident in the committee's third conclusion the committee did acknowledge the possibility that linkages might be useful in some limited circumstances.

Under limited conditions it may be possible to calculate a linkage between two tests, but multiple factors affect the validity of inferences drawn from the linked scores. These factors include the content, format, and margins of error of the tests; the intended and actual uses of the tests; and the consequences attached to the results of the tests. When tests differ on any of these factors, some limited interpretations of the linked results may be defensible while others would not (Feuer, Holland, Green, Bertenthal, & Hemphill, 1999, pp. 91-92).

Little guidance was provided for readers, however, to help distinguish between interpretations that might be defensible from those that would not. The lack of

guidance is understandable given the absence of agreed upon quantitative indices of the degree to which tests are equitable.

In addition to the general comparability question addressed by committee in the *Uncommon Measures* report, the NRC was also asked to address the comparability question within the context of a specific method of linking tests, embedding common items from the NAEP, or another national test, in state tests to construct a common national scale. The NRC appointed a second committee, chaired by Dan Koretz, to address that possibility. Consistent with the *Uncommon Measures* committee, the committee on embedding common items concluded that the embedding approach would not yield comparable scores that were both valid and reliable (Koretz, Bertenthal, & Green, 1999). As in the case of the *Uncommon Measures* report, however, the conclusions in the *Embedding Questions* were reached in the absence of quantitative indices of the degree to which results from linked tests fall short of an equitability criterion.

Some progress has been made since the 1999 publication of *Uncommon Measures* and *Embedding Questions* in quantifying the degree to which linkages of equitable tests differ from those that are not equitable in satisfying the population invariance requirement. Dorans and Holland (2000) proposed two indices for quantifying the degree of invariance and illustrated how these indices vary for tests that can be equated satisfactorily (e.g., two forms of the SAT I Verbal) to those that clearly cannot (e.g., SAT I Verbal with the SAT I Mathematics). They also found the indices took on intermediate values for tests that were similar enough to create concordance tables (e.g., the ACT Mathematics and SAT I Mathematics), but clearly do not satisfy the strict

requirements of equating. More recently, Holland (2005) has provided additional comparisons of the indices for a larger set of linkages of tests ranging from linkages of parallel tests, to tests measuring different constructs (quantitative and language tests) as well as some tests with intermediate degrees of similarity. The indices are sensitive to the degree of similarity of the tests being linked and clearly distinguish between cases where the linkage provides an adequate equating and ones where the linkage does not.

### ***No Child Left Behind***

The No Child Left Behind (NCLB) Act of 2001 has given new salience to questions about the relative stringency of state performance standards and NAEP. Consequently, it may not be surprising that despite the challenges due to frequent observations that linkages among different assessments (e.g., a state assessment and NAEP) may yield results that are invalid to varying degrees and possibly be misleading, there continues to be a demand for additional linkage efforts.

As was noted above, states vary greatly in the percentage of students who score at or above the proficient level on their state assessments. Olsen (2005) reported the percentage of students scoring at the proficient level or above on state assessments in reading and mathematics. The percentages were reported, for example, for 33 states that had grade 8 mathematics assessments and had released data as of the September 7, 2005 publication of *Education Week*. Figure 1 shows a plot of those percentages. The percentage proficient or above ranged from a low of 16% for Missouri to a high of 87% for Tennessee. It is hard to imagine that mathematics achievement in Tennessee is that much better than it is in Missouri or that the definitions of mathematics

achievement in those two states are different enough to explain this tremendous variability. Nor does it seem reasonable to assume that the 63% proficient or above in Alabama or the 53% in Mississippi reflect grade 8 mathematics achievement in those two states that is that much better than the achievement in Missouri or that much worse than the achievement in Tennessee. Rather, it seems more plausible to assume that a large part of the difference is due to differences in the stringency of the performance standards set by the various states.

< Insert Figure 1 about here >

NCLB requires states to administer NAEP reading and mathematics assessments every other year at grades 4 and 8. The second mandatory state NAEP assessments were administered in 2005 and the results were released in the fall of 2005. Although there is no requirement that state assessments be linked to NAEP or any specification of how state NAEP results should be used in considering state NCLB results, NAEP is clearly intended to provide some kind of a rough benchmark against which state results can be judged. In this regard, it is instructive to compare the percentage proficient or above on the 2005 grade 8 mathematics assessment to the results shown in Figure 1 for the individual state assessments. Figure 2 displays the percentages proficient or above on the 2005 grade 8 NAEP mathematics assessment for the 33 states for which state assessment results were displayed in Figure 1.

< Insert Figure 2 about here >

The 25 point (from 13% to 25%) range of percentages in Figure 2 is clearly much smaller than the 71 point range (from 16% to 87%) for the state assessment results in Figure 1. The states with high percentages (e.g.,

Connecticut, Montana, South Dakota and Wisconsin) and low percentages (e.g., Alabama, Mississippi, and New Mexico) in Figure 2 also make more sense in terms of other educational indicators than do the high (Tennessee) and low (Missouri) values in Figure 1. The scatterplot in Figure 3 shows the relationship between the percentage proficient or above on the Grade 8 NAEP mathematics assessment and the grade 8 mathematics assessments used by the 33 states with results reported in the previous two figures. As can be seen in Figure 3, the states are not only much less variable in their percentage proficient or above results on NAEP than on their own state assessments, but the relationship between the percentages is relatively weak. The correlation is only .34, which for state aggregate student performance on achievement tests in the same subject is a rather low correlation<sup>3</sup>.

< Insert Figure 3 about here >

For all 50 states on the 2005 state NAEP assessments, the percentage of students who are proficient or above on the state NAEP grade 4 reading has a correlation of .87 with the percentage proficient or above on the state NAEP grade 8 mathematics assessments. It makes little sense that NAEP grade 4 reading percent proficient would be a better predictor of NAEP grade 8 mathematics percent proficient or above than are the state mathematics assessments in grade 8. The content of the grade 8 NAEP mathematics assessment is surely more similar to the grade 8 state mathematics assessments than the former is to the grade 4 NAEP reading assessment. But there is greater commonality between the reading grade 4 and mathematics

---

<sup>3</sup> It is worth noting that only two of the 33 states have standards that fall above the diagonal line in Figure 3 which represents an equal percentage on the state and NAEP assessments. That is, only two states have standards on their own assessment that are more stringent than the standards on NAEP. A number of points in Figure 3 are well below the diagonal line indicating that the proficient performance standards in those states are substantially more lenient than the NAEP proficient achievement level.

grade 8 achievement levels on NAEP than there is between the state assessment performance standards and the NAEP achievement levels in a single subject and grade. Given the large and obvious differences between the stringency of state performance standards as illustrated by Figures 1 through 3, it seems reasonable to ask whether or not a linking of state assessments with NAEP might not provide a substantial improvement in comparability, despite the fact that a linking clearly would fall short of meeting the criteria for a valid equating.

### ***School-Level State Assessment Score Database***

In the 1998/99 school year a research team at the American Institutes for Research (AIR) launched a project supported by the U.S. Department of Education to assemble a National Longitudinal School-Level State Assessment Score Database (SSASD) (see, for example, McLaughlin, Bandeira de Mello, Cole, Blankenship, Hikawa, Farr, & Gonzalez, 2002). The SSASD has assessment scores for roughly 80,000 public schools in states, the District of Columbia and Puerto Rico for as many as 11 years (1993 thru 2003). The U.S. Department of Education has an interest in using the database to evaluate the effectiveness of federal programs at the school level. Indeed, the database has already been used or proposed for use in the evaluation several programs, e.g. the Reading Excellence Act, the Comprehensive School Reform Demonstration Program, and Title I programs supported under NCLB.

It is obvious that there are substantial challenges to the use of the database to make valid inferences about program effectiveness. Causal inferences are always difficult to justify in the absence of random assignment. Modeling growth and controlling for extraneous variables is never simple and

causal inferences based on the results of such analyses are controversial (see, for example, McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004; Raudenbush, 2004; Rubin, Stuart, & Zanutto, 2004). Other papers in this symposium address these challenges so I will not comment further on them here. My focus is limited to the challenge of making comparisons across the diverse array of assessments used by different states, and sometimes within a state when major changes are made in a state's assessment program.

The state assessment scores available for schools in the SSASD differ in many ways. As was noted above, the assessments differ in content coverage and in format of the assessment tasks. Some states depend primarily on multiple-choice items while others place more reliance on constructed-response items. Although all states make use their assessment results for purposes of NCLB, they differ with respect to the characteristics of state accountability and whether the assessment results have consequences for individual students (e.g., decisions about retention in grade, or high school graduation).

The states also differ in the metrics used for reporting assessment results. All states are required by NCLB to use academic achievement standards (commonly referred to as performance standards) for determining student proficiency and judging adequate yearly progress of schools and districts. As has already been shown, however, there is tremendous state-to-state variability in the relative stringency of the state performance standards. Moreover, states report their assessment results using a variety metrics in addition to the percentage of students who score at or above the proficient level. Percentile ranks, normal-curve equivalents, as well as scale scores with varied metrics are used to report results. Some states use vertical scales

spanning multiple grades while others use within-grade scales with unique metrics.

Given the diversity of state assessments and metrics used for reporting results, program evaluations that rely on results in the SSASD have several options. First, the analyses can be conducted on a state-by-state basis. Alternatively or in addition, state assessment results can be converted to standard scores or effect size statistics and aggregated for an overall analysis as is routinely done in meta analyses. These first two approaches have been used in some program evaluation reports (see, for example, Moss, Gamse, Jacob, Smith, Greene & Kupfer, 2003; Policy and Program Studies Service, 2004). Although using effect size statistics ignoring differences among tests used to compute the statistics is accepted practice in meta analytic studies, it seems reasonable to ask if adjustments for differences among tests would provide an improved basis for analyses. Thus, a third approach would be to adjust for differences among state assessments in the SSASD by linking them to a common assessment such as NAEP.

### ***Using the SSASD to Link State Assessments to NAEP***

Analyses conducted by McLaughlin and Bandiera de Mello (2002) suggest that reasonable linkages between state assessments and NAEP can be obtained using school level data obtained from SSASD. Focusing on schools in the NAEP sample for a given grade and subject, McLaughlin and Bandiera de Mello related the percentage of students meeting a standard on the state assessment to the NAEP scale score above which an equal fraction of the students in the school achieved. This process was repeated for the schools in the NAEP sample for a given state. The result was a set of estimates of the

NAEP scale score corresponding to each performance standard on the state assessment. Variability in the estimates obtained from different schools was used to compute a standard error of estimate for each translation of a state performance standard into the NAEP scale.

McLaughlin and Bandeira de Mello (2002) used the approach just described to link state assessment data to NAEP data for 29 states that participated in the 2000 grade 4 NAEP mathematics assessment. They used the linkages to compare the performance standards set on state assessments to the NAEP performance standards (achievement levels). According to their analyses, the highest performance standard identified by the 29 states was never as stringent as the NAEP advanced level, though Maine's highest level, called "exceeds standard," was nearly on a par with the NAEP advanced level.

The standard errors associated with the estimated translation to the NAEP scale were generally small to modest (i.e., 12 NAEP scale points or less – about a third of a standard deviation) for the state proficient standard and for the highest standard in a state (usually called advanced). For the lower state standards (e.g., basic, near proficient, emerging, inconsistent mastery), however, the standard errors were substantially larger, generally ranging from about 15 to 25 NAEP scale points and in the most extreme case the standard error was 35 for the lowest standard in Massachusetts. McLaughlin and Bandeira de Mello (2002) speculated that the larger standard errors for the lower level performance standards may be partially the result lower reliability on NAEP at the lower end of the scale.

In a follow-up paper, McLaughlin and Bandiera de Mello (2003) explored the idea of using NAEP to confirm gains that states report on their own

assessments. Using the same approach of matching the percent of students meeting the state performance standard in each school to the distribution of plausible values in the school's NAEP sample, McLaughlin and Bandiera de Mello (2003) provided translations of state performance standards to the 2002 NAEP reading at grades 4 and 8. As was found for mathematics, the links for reading were stronger (i.e., smaller standard errors) at the proficient and advanced state performance standards levels than for lower performance standards.

Braun and Qian (2005) have reported analyses that make use of NAEP's sampling weights in and NAEP's jackknife procedure in their calculations. They referred to their method as "weighted aggregate mapping" and called the McLaughlin and Bandiera de Mello method "unweighted local mapping." These technical refinements of the McLaughlin and Bandiera de Mello (2002, 2003) method arguably may provide the basis for better confidence intervals for the NAEP scale scores corresponding to the performance standards on state assessments (see, also, Thissen, 2005).

Braun and Qian applied their method to link 2000 state mathematics assessments to NAEP mathematics assessments in 2000 at grades 4 and 8 and to link 2002 state reading assessments to the 2002 NAEP reading assessments. They also compared their weighted results to the unweighted results using the McLaughlin and Bandiera de Mello method. The pattern of results was found to be quite similar for the weighted aggregate mapping and the unweighted local mapping methods, but Braun and Qian argued that the weighted aggregate mapping has conceptual advantages. To support the reasonableness of their results, Braun and Qian investigated the relationship

between the proportions of students meeting standards in different states and the estimated NAEP equivalent scores. For both mathematics and reading they found a strong negative relationship between the proportions of students meeting state standards and the stringency of the state standards as indicated by the NAEP equivalent scores. Koretz (2005), however, has argued that the strong negative correlations are to be expected and do not provide evidence that the mapping is reasonable, noting that the correlation would be perfect if there was census testing on NAEP and an equipercentile linking was used.

Similar to the findings of McLaughlin and Bandeira de Mello (2002), Braun and Qian found that states varied widely in the linked NAEP scale scores corresponding to their proficient standards. The 2000 grade 4 proficient standard in Louisiana, for example was mapped to a NAEP scale score of 250.8 while the pass standard, which corresponded to the proficient standard in Texas in 2000, mapped to a NAEP scale score of 200.6 (Braun & Qian, 2005, Table 3.2, pp. 34-25 of typescript)<sup>4</sup>. A difference of 50 NAEP scale score points is greater than a standard deviation, and suggests that some type of adjustment needs to be made if results on the different state assessments if the results are to be compared. The .95 confidence intervals around the two point estimates (248.3 to 253.3 for Louisiana and 197.5 to 203.7 for Texas) do not even come close to overlapping. Thus, there would seem to be strong support that in 2000 the proficient performance standard for grade 4 mathematics was a good deal more stringent in Louisiana than it was in Texas.

It is clear that the SSASD provides a means of linking state performance standards to the NAEP score scale, and therefore to the NAEP achievement

---

<sup>4</sup> It should be noted that Texas has revised their state assessments and set new standards since 2000.

levels. The NAEP equivalence scores are subject to a good deal of uncertainty, however. Because of the uncertainty, Braun and Qian (2005) cautioned against using their mapping to make fine distinctions and propose that confidence intervals accompany reports of the NAEP equivalent scores. The linkages, with or without incorporating confidence intervals, make it possible to conduct analyses with data aggregated across states on a common metric. It is less clear, however, that such analyses would have an advantage over the current practice of conducting separate analyses within each state and using standard scores or effect size statistics for analyses across states.

### ***Conclusion***

An equivalency scale on which results on different tests could be reported and validly compared would have considerable utility. Such a scale would greatly simplify comparisons and facilitate aggregate analyses of results from different tests. The demand for such a scale has a history that spans at least four decades. There have been numerous explorations of the possibility which led to conclusions about shortcomings of linkages obtained. Despite discouraging results and cautions of many experts on equating, the demand for finding a way to compare results of diverse tests and treat the results as if they were interchangeable seems to have just grown stronger rather than weaker over the last 40 years.

Although NAEP would seem to hold the greatest promise for linking state assessments, Thissen (2005) has recently questioned whether differences in student motivation when taking NAEP and when taking their state assessments may seriously undermine the quality of linkages between NAEP and state assessments. NAEP has no consequences either for individual students or for

the schools they attend. State assessments, on the other hand, clearly have important consequences for schools and in a number of states they also have significant consequences for individual students. Thissen (2005) summarized his concern about possible effects of differences in motivation as follows.

A rhetorical question that could haunt attempts to estimate proportions scoring on NAEP's (low stakes) achievement levels from (high stakes) statewide assessment is "How many students would be classified as 'proficient' by NAEP if they thought, when NAEP is administered, that it was their statewide assessment and that the[y] might be retained in grade if they didn't answer enough questions correctly?" (p. 34 of typescript).

The possible effects of differences in student motivation when taking NAEP than when taking a state assessment can not only influence the linkage a given point in time, but in may also result in a divergence in trend lines. The lack of invariance of the linkages over time may also be the result of instruction that is closely geared to the idiosyncratic nature of item formats on the state assessment, leading to gains in performance that do not generalize to other measures. Koretz (2005) has presented a detailed discussion of the latter issue, and argued that both students and teachers have strong incentives to behave in ways that inflate scores on state assessments.

It is easy to see that the strict requirements of equating are unlikely to be met for assessments that are not specifically designed to be interchangeable. It is also easy to identify the ways in which linkages between non-comparable assessments will fall short of the equity and population invariance requirements for equating. We need more work along the lines of Dorans and Holland

(2000), (see also Holland, 2005), however, to provide a quantitative basis for determining how far short linkages between tests fall of the ideal for equating. A better understanding is also needed of what inferences can justifiably be made from linkages of non-equivalent tests and inferences from such linkages that cannot be justified.

## References

- Angoff, W. H. (1964). Technical problems of obtaining equivalent scores on tests. *Journal of Educational Measurement*, 1, 11-13.
- Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement, 2<sup>nd</sup> Edition* (pp. 508-600), Washington, DC: American Council on Education.
- Bianchini, J. C., & Loret, P. G. (1974). *Anchor test study. Final report. Project report and volumes 1 through 30, and Anchor test study supplement. Final report. Volumes 31 through 33.* (ERIC Documents Reproduction Service Numbers ED 092 601 through ED 092 634).
- Braun, H. I. and Qian, J. (2005), *Mapping state performance standards onto the NAEP scale.* Paper presented at the ETS Conference, Linking and Aligning Scores and Scales: A Conference in Honor of Ledyard R Tucker's Approach to Theory and Practice, Princeton, NJ: Educational Testing Service, June 24.
- Dorans, N. & Holland, P. W. (2000). Population invariance and the equitability of tests: Basic theory and the linear case. *Journal of Educational Measurement*, 37, 281-306.
- Elementary and Secondary Education Act of 1965, Pub. Law No. 89.10.
- Ercikan, K. (1997). Linking statewide tests to the National Assessment of Educational Progress: Accuracy of combining test results across states, *Applied Measurement in Education*, 10, 145-159.
- Feuer, M. J. (2005). E Pluribus Unium: Linking tests and democratic education. In C. A. Dwyer (Ed.), *Measurement and research in the accountability era* (pp. 165-183). Mahwah, NJ: Lawrence Erlbaum.

- Feuer, M. J. (In press). *Better than best: A rational approach to education reform*. Cambridge, MA: Harvard University Press.
- Feuer, M. J., Holland, P. W., Green, B. F., Bertenthal, M. W., & Hemphill, F. C. (1999). *Uncommon measures: Equivalence and linkage among educational tests*. Washington, DC: National Academy Press.
- Flanagan, J. C. (1964). Obtaining useful comparable scores for non-parallel tests and test batteries. *Journal of Educational Measurement*, 1, 1-4.
- Goals 2000: Educate America Act of 1994, Public Law 103-227, Sec. 1 et seq. 108 Stat. 125 (1994).
- Holland, P. W. (2005). Assessing the validity of test linking. In C. A. Dwyer (Ed.), *Measurement and research in the accountability era* (pp. 185-195). Mahwah, NJ: Lawrence Erlbaum.
- Improving America's Schools Act of 1994, Public Law 103-382, Sec. 1 et seq. 108 Stat 35424 (1994).
- Koretz, D. (2005). *Using aggregate-level linkages for estimation and validation: Comments on Thissen and Braun & Qian*. Paper presented at the ETS Conference, Linking and Aligning Scores and Scales: A Conference in Honor of Ledyard R Tucker's Approach to Theory and Practice, Princeton, NJ: Educational Testing Service, June 24.
- Koretz, D. M., Bertenthal, M. W., & Green, B. F. (1999). *Embedding questions: The pursuit of a common measure in uncommon tests*. Washington, DC: National Academy Press.
- Lennon, R. T. (1964). Equating non-parallel tests. *Journal of Educational Measurement*, 1, 15-18.

- Lindquist, E. F. (1964). Equating scores on non-parallel tests. *Journal of Educational Measurement*, 1, 5-9.
- Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education*, 6, 83-102.
- Linn, R. L. & Kiplinger, V. L. (1995). Linking statewide test to the National Assessment of Educational Progress: Stability of results, *Applied Measurement in Education*, 8, 135-155.
- Linn, R. L. & Slind, J. A. (1977). The determination of the significance of change between pre and posttesting periods. *Review of Educational Research*, 47, 121-150.
- McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational Statistics*, 29, 67-101.
- McLaughlin, D. H. (1998). *Study of linkages of 1996 NAEP and state mathematics assessments in four states*. Washington, DC: National Center for Education Statistics.
- McLaughlin, D., Bandeira de Mello, V. (2002). *Comparison of state elementary school mathematics achievement standards, using NAEP 2000*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, April.
- McLaughlin, D., Bandeira de Mello, V. (2003). *Comparing state reading and math performance standards using NAEP*. Paper presented at the CCSSO National Conference on Large-Scale Assessment, San Antonio, June.
- McLaughlin, D., Bandeira de Mello, V., Cole, S., Blankenship, C., Hikawa, H., Farr, K., & Gonzalez, R. (2002). *National longitudinal school-level state*

*assessment score database: Analyses of 2000/2001 school year scores.*

Palo Alto, CA: American Institutes for Research.

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

Moss, M., Gamse, B., Jacob, R., Smith, W. C., Greene, D., & Kupfer, A. (2003). *Reading excellence act and school implementation and impact study: Annual report 2002-2003.* Cambridge, MA: Apt Associates Inc.

No Child Left Behind Act of 2001, Pub. Law No. 107.110.

Olsen, L. (2005). Defying predictions, state trends prove mixed on schools making NCLB targets, *Education Week*, 25(2), September 7, pp. 1, 26-27.

Peterson, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational Measurement*, 3<sup>rd</sup> Edition (pp. 221-262), New York: Macmillan.

Policy and Program Studies Service. (2004). *Implementation and early outcomes of the comprehensive school reform demonstration (CSR) program.* Washington, DC: U.S. Department of Education, Doc # 2004-15

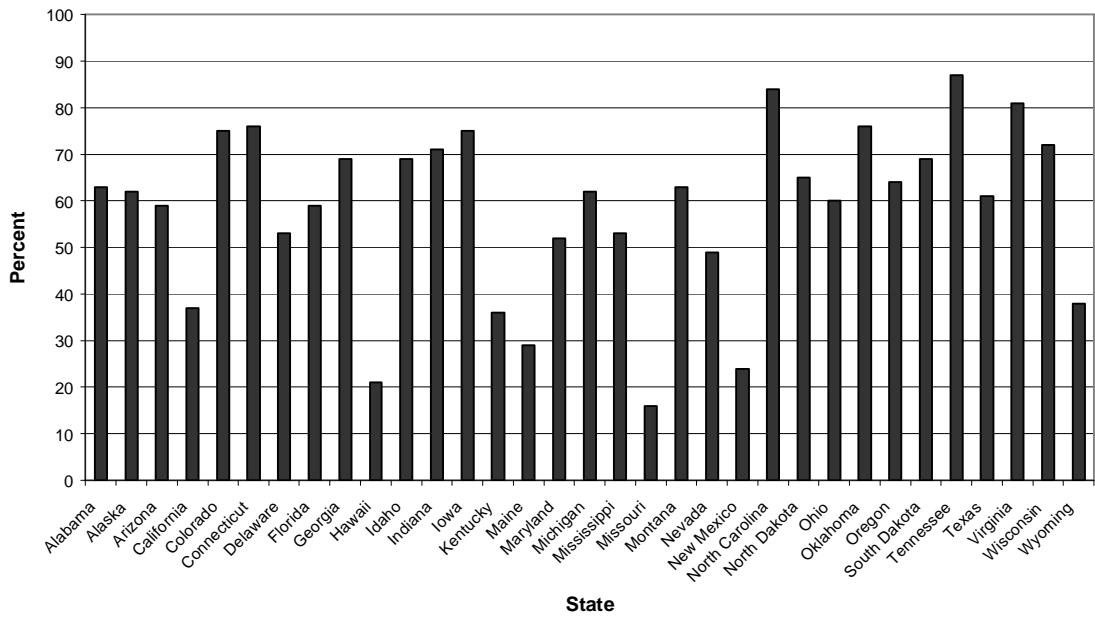
Raudenbush, S. W. (2004). What are value-added models estimating and what does this imply for statistical practice? *Journal of Educational Statistics*, 29, 121-129.

Rubin, D. B., Stuart, E. A., & Zanutto, E. L. (2004). A potential outcomes view of value-added assessment in education. *Journal of Educational Statistics*, 29, 103-116.

Tallmadge, G. K. & Wood, C. T. (1981). *User's Guide: ESEA Title I evaluation and reporting system.* Mountain View, CA: RMC Corporation.

- Thissen, D. (2005). Linking assessments base on aggregate reporting: Background and Issues. Paper presented at the ETS Conference, Linking and Aligning Scores and Scales: A Conference in Honor of Ledyard R Tucker's Approach to Theory and Practice, Princeton, NJ: Educational Testing Service, June 24.
- Waltman, K. K. (1997). Using performance standards to link statewide achievement results to NAEP, *Journal of Educational Measurement*, 34, 101-121.
- Williams, V. S. L., McLeod, L. D., Thissen, D., & Sanford, E. (1988). Projecting to the NAEP scale: Results from the North Carolina End of Grade testing program, *Journal of Educational Measurement*, 35, 277-296.

**Figure 1**  
**Percent Proficient or Above on State Mathematics Assessments in 2005 (33 states, Source Ed Week)**



**Figure 2**  
**Percent Proficient or Above on 2005 Grade 8 NAEP Mathematics Assessment for**  
**33 States with Grade 8 State Assessment Results Reported by Ed Week**

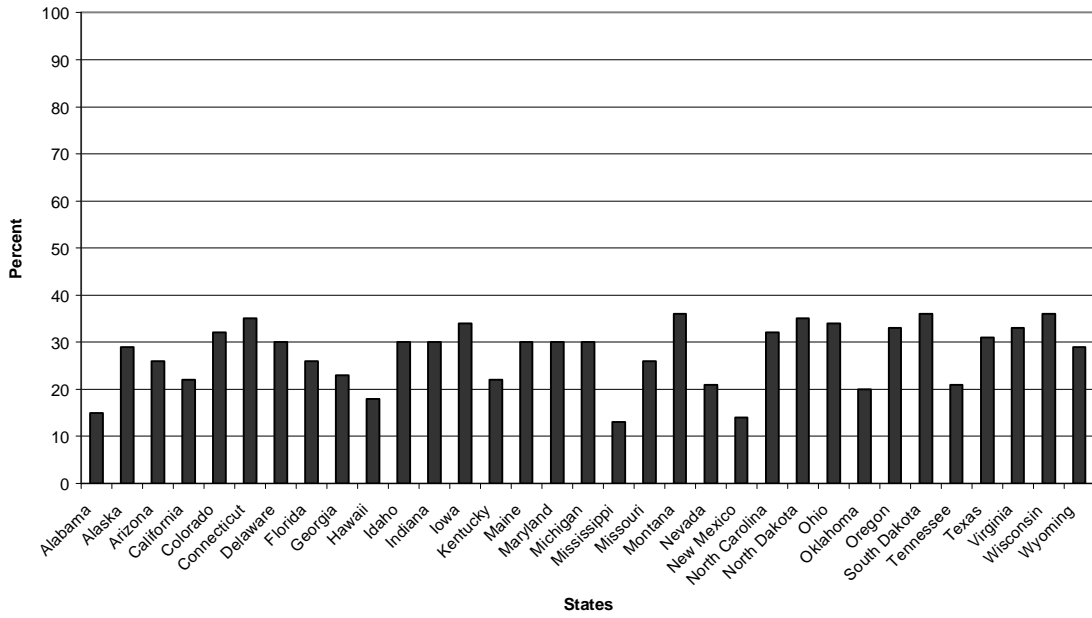


Figure 3  
Scatter Plot of Percent Proficient or Above on Grade 8 State Mathematics Assessments and  
Grade 8 NAEP in 2005 for 33 States ( $r = .34$ )

