Utility and Validity of NAEP Linking Efforts

Robert L. Linn
Center for Research on Evaluation, Standards, and Student Testing
University of Colorado at Boulder

Donald McLaughlin
Statistics and Strategies

David Thissen
L. L. Thurstone Psychometric Laboratory
University of North Carolina at Chapel Hill

September 2009
Commissioned by the NAEP Validity Studies (NVS) Panel

George W. Bohrnstedt, Panel Chair
Frances B. Stancavage, Project Director

The NAEP Validity Studies Panel was formed by the American Institutes for Research under contract with the National Center for Education Statistics. Points of view or opinions expressed in this paper do not necessarily represent the official positions of the U.S. Department of Education or the American Institutes for Research.

The NAEP Validity Studies (NVS) Panel was formed in 1995 to provide a technical review of NAEP plans and products and to identify technical concerns and promising techniques worthy of further study and research. The members of the panel have been charged with writing focused studies and issue papers on the most salient of the identified issues.

Panel Members:

Albert E. Beaton
Boston College

Peter Behuniak
University of Connecticut

George W. Bohrnstedt
American Institutes for Research

James R. Chromy
Research Triangle Institute

Phil Daro
University of California, Berkeley

Lizanne DeStefano
University of Illinois

Richard P. Durán
University of California, Santa Barbara

David Grissmer
University of Virginia

Larry Hedges
Northwestern University

Gerunda Hughes
Howard University

Robert Linn
University of Colorado at Boulder

Donald M. McLaughlin
Statistics and Strategies

Ina V.S. Mullis
Boston College

Jeffrey Nellhaus
Massachusetts State Department of Education

P. David Pearson
University of California, Berkeley

Lorrie Shepard
University of Colorado at Boulder

David Thissen
University of North Carolina, Chapel Hill

Project Director:

Frances B. Stancavage
American Institutes for Research

Project Officer:

Janis Brown
National Center for Education Statistics

For Information:

NAEP Validity Studies (NVS)
American Institutes for Research
1070 Arastradero Road, Suite 200
Palo Alto, CA 94304-1334
Phone: 650/ 843-8192
Fax: 650/ 858-0958

Acknowledgments

The authors are indebted to the NVS panel members and to Fran Stancavage, the NVS project director, for their helpful comments during the development of this white paper.

Contents

List of Tables

List of Figures

Introduction

There are a number of practical situations in which it would be desirable to be able to use the results of the administration of one assessment to estimate what the results would have been if another assessment had been administered. Test linking refers to the idea that results obtained from the administration of one test might be used to infer what the results would have been if another test had been used. Common knowledge, based on widespread experience with educational testing in the American culture, provides two contradictory views of the utility of test linking.

One view is that test linking is obviously easy and useful, because tests of the same subject are more or less interchangeable: "a mathematics test is a mathematics test." That is, if one gives a mathematics test to a group of students, one should be able to infer from the results something about what the results would be if one had given a different mathematics test; either test yields scores that indicate the amount or level of mathematics the students know or can do, because "amount of mathematics" is an aspect of the students, not the particular test.

Both implicit and explicit uses of this view of testing abound. For example, it is the implicit basis of the reliance on educational assessment that is a key feature of *No Child Left Behind*; a central idea of that program is that the proportion of students who obtains scores above a cut score on a particular test are "proficient"—there is a certain amount of mathematics they know or can do. The idea, embodied in legislation, is that many specific properties of the test (and the cut scores) do not affect this inference. A more explicit use of the idea that different tests can produce interchangeable scores involves the widely known fact that most large-scale tests (e.g., statewide assessments, college entrance tests) use multiple "forms" comprising different items, yet we trust those testing programs to provide comparable scores from such different forms.

A polar opposite view, also supported by widespread experience with educational testing, is that the test content (what questions are asked), the conditions under which they are administered, and the cut scores that distinguish proficient and below-proficient achievement make a great deal of difference in the results obtained when tests are administered. Students complain that "the test was too hard," meaning they would appear to have done better if different (easier) questions had been asked. Where statewide testing programs use alternate test forms, critics complain that it is "not fair" that some students answer one set of questions while others answer a different set of questions. Data analysts at the district or school level may expend substantial energy looking for differences between results obtained with different test forms, to validate a perception that "the red form was easier," and to explain undesired assessment results. Put simply, this view is that everyone knows that many factors affect assessment results, and one cannot know what results would have been obtained with a test that was not administered, based on the scores on a different test.

Both of these completely opposite views are correct, to a point. Many subtle distinctions are made in psychometrics to create a systematic understanding of the various statistical methods that can be used to answer varying questions in different contexts, many of which fall under the omnibus denotation "test linking." The purpose of this essay is to summarize that understanding, with illustrations of continua of contexts, questions, and statistical methods, and with results that range from the production of comparable scores from different tests (satisfying the first view) to a conclusion that different tests yield different results (the second, opposite, view).

Test Linking: Distinctions and Nomenclature

Holland (2007) has provided a framework for thinking about different categories of linking that builds on earlier work he did with Dorans (Holland & Dorans, 2006) as well as the work of Mislevy (1992), Linn (1993), Kolen (2004), and Kolen and Brennan (2004). At the broadest level, Holland (2007) distinguishes three types of links between tests *X* and *Y*. These are (1) *equating* *X* and *Y*, (2) *aligning the scales* of *X* and *Y*, and (3) *predicting* *Y* from *X*. Table 1 summarizes these three broad categories of linkages, categorizing them under the contexts in which each is done, and providing examples of the questions each answers.[1]

[1] Some of the lower level detail of Holland and Dorans' (2006) categorization that is not relevant for the present discussion is omitted in table 1.

Table 1. Contexts, Questions, Methods, and Requirements or Limitations for Describing the Relations Between Test-Score Scales

Context: “Forms” of tests constructed based on the same specifications.

Question: Can one produce scores from those various forms that are in all respects interchangeable?

Nomenclature	Methods	Requirements
Test Equating	Many distribution-matching procedures (see Holland and Dorans, 2006)	(a) *Equal Construct*: Tests should measure the same constructs. (b) *Equal Reliability*: Tests should have the same reliability. (c) *Symmetry*: 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) *Equity*: 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*: 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 (Holland and Dorans, 2006, p. 194).

Context: “Forms” of tests, or entire tests, constructed based on specifications that differ in some respects.

Question: Can one produce scores from those various forms that are in some sense(s) comparable?

Nomenclature	Methods	Known *a priori* differences between tests/forms	Limitations
Scale Aligning: Calibration Vertical Scaling	Item response theory (IRT) models most often, or others	Forms measure the same construct, but with different reliability, and/or different difficulty	(b) is known to be false *a priori*; (d) is unlikely to be entirely true as a result. (e) is always an empirical question.
Concordance	Equating-like methods	Forms measure similar, but not identical, constructs.	(a) is violated to some degree; consequent violation of (e) may provide empirical evidence of the usefulness of the linking; (d) may be false.

Context: Different tests that should be related in some way.

Question: Can one predict scores or other outcomes on one test from results obtained with the other?

Nomenclature	Methods	Known *a priori* differences between tests/forms	Limitations
Prediction	Regression	Tests measure different constructs, probably with different difficulty, and perhaps with different reliability.	(a) and (c) are not true. (e) has most often been found to be empirically false, which means (d) is not true. (b) depends on the tests, but is largely irrelevant given that (a) and (c) through (e) are false.
Projection	Regression with distributional reconstruction		

Context: Different tests that should be related in some way.

Question: Can one describe the relations between scores obtained with one test and those obtained with the other?

Nomenclature	Methods	Known *a priori* differences between tests/forms	Limitations
Analysis of the “nomothetic span” (Embretson, 1983) of a test—an aspect of validity analysis	Many data analytic methods, but most often the same regression methods as for prediction or projection	Tests measure different constructs, probably with different difficulty, and perhaps with different reliability.	There are no real limitations, because there is no *a priori* requirement of a claim that a result will be scores that are interchangeable or even comparable. There are results that describe the relations between the tests.

A prototype of test linking is *equating*. The context is that one has alternate forms of a test; each form is constructed following the same rules based on the same framework and specifications. The goal is to provide scores, usually on a widely known reporting scale, that are in all respects interchangeable. Examples would be alternate forms of college admissions tests, such as the SAT or ACT. Holland and Dorans (2006) identify five requirements for equating two parallel tests that enjoy broad professional consensus; they are summarized in table 1. The requirements are demanding, and they are likely to be met only in the specified context (that each form is constructed in the same way). Specifically, requirement (a), that the forms measure the same constructs, is guaranteed, not by data analysis, but rather by the common method of form assembly. Requirement (b) (equal reliability) can be checked empirically, but generally follows from the identical construction of the forms. Requirement (c), symmetry, is a technical aspect of the procedures used to construct the equating relation. Requirement (d), equity, for the most part follows from meeting requirements (a) through (c) and (e). Requirement (e), population invariance, is listed because it provides the most accessible empirical check on requirement (a): If the relation between two test forms differs between identifiable (sub)populations, that must mean that somehow the two forms measure something different. If they measure something different, their scores cannot be used interchangeably; (d), equity, could be violated.

When the word *equating* is used with its strictest meaning, it is not a very general procedure. It applies only to alternate forms of the same test. There are closely related categories of procedures that Holland and Dorans (2006) refer to as *scale aligning* that are more general. The context involves tests that have specifications that differ in some respects. In one sense, the most minor difference between two tests would be that they differ in reliability or difficulty. For example, a shorter form of a test would be less reliable than a longer form, but otherwise the same; a fourth-grade reading test might be more difficult than a third-grade reading test, but otherwise the same. *Calibration* or *vertical scaling* provides comparable scores in such contexts. These contexts may (obviously) violate requirement (b) of equating (equal reliability), and usually violate requirement (d) (equity), because examinees may prefer the version of the test that provided them the best chance of passing, and which version this is may differ for different examinees. On the other hand, the degree of comparability provided by scale alignment may be sufficient for many purposes.

The creation of a *concordance* uses the statistical methods of equating to match scores on tests that do not meet the requirements for equating. The context makes the difference; for concordance, the tests are constructed using different frameworks and specifications. The most well-known example involves concordance tables for ACT and SAT scores that are used in college admissions. These tables are made not so much because they are accurate predictions of the score that would be obtained on one test, given the score on the other, as because they are requested by users (Dorans, 2004; Pommerich, Hanson, Harris, and Sconing, 2004). They produce scores that appear to be comparable, but in many respects are not. For example, group differences (like the difference between males' and females' average scores) may be different between the real test score averages and the averages obtained by translating scores earned on the other test through the concordance.

The contexts most important for the current paper involve tests that are "related in some way." In the categorization of linking methods by Holland and Dorans (2006), procedures to answer the question "Can one predict scores or other outcomes on one test from results obtained with the other?" are answered with *prediction* methods (if one is looking for a point estimate) or *projection* (if one is characterizing a distribution of test scores). The fact that, in this context, the relation between the two tests often varies among (sub)populations makes this kind of linkage of uncertain value: If the linkage varies among (sub)populations, then different aggregated populations have different aggregated linkages. That is, the results change if the "mix" of the population changes. That does not provide the monolithic "answer" usually expected of test linking.

For the purpose of discussion in this essay, near the bottom of table 1 we add one more question that takes us from the context of test linking to the broader context of validity research: "Can one describe the relations between scores obtained with one test and those obtained with the other?" One can always create such descriptions; indeed, that is the kind of validity research that Embretson (1983) referred to as providing a description of the *nomothetic span* of a test. This may be the question that is really being asked when a "linkage" is requested between two different tests. If this is the question, then research describing the relation between scores on the two tests cannot "fail." For example, a lack of population invariance in the relation between two tests may be a reason that a projection-based linking is said to "fail" (because it cannot be counted on to give accurate overall results across aggregated populations). However, if that research had been intended to *describe* the relation between scores on the two tests, then the lack of population invariance is simply part of the results.

Another consideration that will arise in the remainder of this paper is that some uses of test linking, which focus on aggregate scores, can withstand greater threats to accuracy than can uses which directly affect individuals' opportunities for educational and career advancement. Only in equating is error generally smaller than the "points" on the scale on which a test score is reported. But for uses other than individual decision-making, more approximate results describing the relations between test scores may be useful.

Historical Hesitance to Link Different Tests

Interest in estimating scores on one assessment from those on another arises in many contexts and is not just of recent origin. In 1964, for example, a symposium entitled "Equating Non-Parallel Test Scores" was held at the 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.*

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 that would let users substitute scores from one test for those of another. 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. That is, they would not satisfy requirement (e), population invariance, in table 1.

It would appear, however, that the reservations of these leaders in the field of educational measurement did little 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 and to expand into many uses beyond those under consideration in 1964.

Concordance Tables for College Admissions Tests

One of the most commonly used linkages between two tests that measure different constructs is used to create concordance tables between the SAT and ACT.[2] The ACT comprises multiple-choice tests that cover four skill areas (English, mathematics, reading, and science). The scores for each of the four tests and the composite, which is an average of the four scores rounded to the nearest integer, are reported on a scale that ranges from 1 to 36. The ACT also offers an optional writing test that is not included in the ACT composite score. The SAT currently comprises three tests (critical reading, mathematics, and writing). Each of the SAT tests is reported on a scale that ranges from 200 to 800.

ACT-SAT concordance tables are intended to serve several purposes. They are intended to provide students who have taken both tests with a way of comparing their scores on the two tests to see whether they earned a relatively higher score on the ACT or the SAT. They are intended to provide students who have taken either the SAT or the ACT with a basis for estimating what their score would have been if they had taken the other test. They are also intended to be useful to colleges that have some applicants who have taken the SAT and other applicants who have taken the ACT. And they may be useful to counselors in advising students.

Concordance tables have been jointly constructed by the sponsors of the SAT and the ACT: the College Board and the ACT. The tables that receive the greatest emphasis report the concordance between the ACT composite, or ACT sum score, and the sum of the SAT critical reading and mathematics tests (or for earlier versions of the SAT, the sum of the verbal and mathematics tests). In addition, separate concordance tables have been constructed for the ACT and SAT mathematics tests, and for the ACT and SAT writing tests. A concordance table has also been developed for the ACT composite and the sum across all three SAT tests. The latter concordance is not reported by the College Board because of its belief that the fact that there is no essay included in the ACT composite, while there is an essay component of the SAT writing test, makes the content of the three-part SAT and the four-part ACT too dissimilar for a concordance.

Descriptions of concordance tables, along with cautions about interpretation and use of those tables, are provided by the College Board and ACT. ACT, for example, notes that the tests measure "similar but distinct constructs" (www.act.org/aap/concordance). In a similar

[2] Dorans (2004, p. 244) takes the position that two similar tests (ACT Math and SAT I Math, for example) "measure the same construct" but are built to different specifications. He goes on to say that "the construct … is not a property of the test. The construct is a characteristic of the examinees …" As used in this essay, a construct *is* a property of the test—the construct is defined by the pattern of covariation among responses to items on the test. This yields differences in wording between this essay and Dorans (2004), but the difference is one of philosophy of science as it relates to constructs, not one of substance.

vein, the College Board states that "it is impossible to predict exactly what score a student will get on one test, based solely on the score obtained on the other test" (http://professionals.collegeboard.com/data-reports-research/sat/sat-act). Nonetheless, the College Board and the ACT have worked together to produce concordance tables linking the two tests on a number of occasions during the past 30 years because they recognize that that such tables provide useful approximations for college admissions officers and counselors as well as students and their parents.

Although the ACT and the College Board have obviously concluded that their concordance tables have sufficient utility for the organizations to produce them, they present the following cautions with the tables.

> Concordance tables are dependent upon the sample used to establish the relationship between the two sets of scores. The ACT-SAT tables are based on an entire cohort of students who completed both tests, but this sample is not representative of either all ACT or SAT test-takers. The tables therefore may not be appropriate for use with scores from students who take either the ACT only or SAT only. Overall, a student who receives a score on one test will not necessarily obtain the concorded score on the other test" (concordance tables, p. 2).

Despite the caution that the conversion may not be appropriate for use with students who have taken only one of the two tests, it is, of course, precisely those students for whom there is considerable interest in the concordance table results. There is no need to estimate what the score of a college applicant who took the ACT would have been if he or she had taken the SAT when, in fact, the student has taken both tests.

The sum of the SAT critical reading and mathematics scores is on a scale of 400 to 1600 and is reported in 10 point intervals for a total of 121 possible scale score points, whereas the ACT composite is reported on a scale from 1 to 36. The difference in number of possible scale score points means that several possible SAT sum scores may be concorded to a single ACT scale score. For example, SAT sum scores of 1540 through 1590 are mapped into an ACT composite score of 35 and sum scores of 670 through 710 are mapped into an ACT composite score of 14 (http://www/act.org/aap/concordance/index.html).

Because a concordance of the SAT and the ACT does not satisfy the requirements of equating, it is not surprising that the results may differ when different populations of students are used to construct concordance tables. Sawyer (2007) summarized results reported earlier by Dorans, Lyu, Pommerich, & Houston (1997) that showed that concordances based on student results from two different colleges varied substantially. Given an ACT score, the concorded SAT sum scores (V+Q) varied by as much as 40 points between the results that would have been obtained had the concordance been based on applicants to one of the colleges or the other. This is a concrete example of the meaning of a lack of invariance between groups (in this case, between the applicant pools for the two colleges).

Of course, an answer to the question of the importance of that 40-point difference depends on the use that would be made of the linking (concordance). Clearly if the linked scores were being used to compute the average level of applicants on the SAT scale (using their ACT scores as the basis of the inference), many would consider a result that could be as much as 40 points off to be substantially in error. On the other hand, Sawyer (2007) points out that when consistency rates were computed for the classification of applicants near the admission cutoff scores for those two schools using a school-specific concordance, rather than the national average table, the change in consistency was only about 0.01. This result is obtained because other error components of the scores used for individual classification are of the same order of magnitude, and may be larger, than even this linking error.

This contrast between less and more useful applications of concordance illustrates the observation that some uses of test linking can withstand greater threats to accuracy than others.

Comparisons of Aggregate Results

The ACT-SAT concordance tables are intended for use in the interpretation of scores for individual students. For state and national assessments of elementary and secondary students, the comparisons across different assessments, on the other hand, are focused on *distributions* of scores. Interest has expanded in recent years in using scores on one assessment to estimate what the distribution of scores on another assessment would have been if the second assessment had been administered.

Interest has grown in situations where results for groups taking different assessments, e.g., states or nations, are to be compared. For example, there is considerable interest in being able to estimate from the performance of students on a state assessment what the performance on the National Assessment of Educational Progress (NAEP) for students in that state would have been had NAEP been administered. In a similar vein there is interest in being able to estimate how students in another country would have performed on NAEP based on their performance on an international assessment such as the Trends in International Mathematics and Science Study (TIMSS) or the Programme for International Student Assessment (PISA). A question has also been raised about whether results to be obtained from the next High School Longitudinal Study (HSLS:09) might be made more interpretable by using a linkage to the NAEP scale. Additional historical background is useful when considering these possibilities.

The interest in comparing achievement of U.S. students and students in the various states with that of students in other nations is illustrated by a report released in December 2008 by the International Benchmarking Advisory Group formed by the National Governors Association (NGA), the Council of Chief State School Officers (CCSSO), and Achieve, Inc. The report, entitled "Benchmarking for Success: Ensuring U.S. Students Receive World-Class Education," called for the benchmarking of state standards, curricula, and assessments to international standards, and emphasized the need to draw on lessons from high-performing nations and states (NGA, CCSSO, & Achieve, Inc., 2008). There is also considerable interest at the U.S. Department of Education in the idea of international benchmarking as a way of ratcheting up state standards and improving student achievement.

Standards-Based Assessments

The switch from the use of norm-referenced tests to standards-based assessments by states in the 1990s contributed to the interest in techniques that would allow comparisons across different assessments. While the half dozen or so different norm-referenced tests used by states in the 1980s and early 1990s differed in content coverage and in the degree to which their norms were representative, they at least gave the appearance of providing a basis of comparing state results to national results. The state assessments introduced in response to the 1994 *Goals 2000: Educate America Act* and the reauthorization of the 1965 *Elementary and Secondary School Act (ESEA)* by the *Improving America's Schools Act (IASA)* of 1994 were more variable in content coverage and format than the norm-referenced tests that states had previously used. They also relied on performance standards rather than norms for reporting results, and those performance standards could not be considered comparable from state to state.

Individuals wanting to compare scores on tests used by different states were faced with a relatively unique assessment in each state, rather than the handful of most commonly used norm-referenced achievement tests. 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 who felt 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, 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 was 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 a single equivalency scale.

Links to State Assessments

Despite the negative conclusion of the NRC committee, there is a keen interest in being able to link state assessments to NAEP. As was noted above, states have introduced a variety of different and non-comparable statewide tests. The expansion of NAEP in 1990 to include administrations of NAEP at the state level raised the possibility of using NAEP as a common yardstick for comparing the results of different state tests.

Several efforts were made to link state tests to NAEP during the early 1990s. The Kentucky Department of Education, for example, attempted to link the tests used in the Kentucky Instructional Results Information System (KIRIS) to NAEP. The link between KIRIS and NAEP was intended to provide a means of judging whether gains made by Kentucky students on KIRIS were reflected by changes in performance on NAEP. The fact that gains on KIRIS were considerably larger than gains on NAEP led to the conclusion that the gains on KIRIS were due more to inflation of scores than to real changes in student achievement (Hambleton, Jaeger, Koretz, Linn, Millman, & Phillips, 1995; Koretz & Barron, 1998). Whatever the cause for the divergence of results, it is clear that the linking of KIRIS and NAEP was not invariant over time. For a more complete description of the KIRIS-NAEP linking, see Thissen (2007).

Ercikan (1997) conducted equipercentile linkings of the 1990 state NAEP results and the state tests in four states that used tests developed by CTB-McGraw-Hill. She found that the linkings differed substantially among the four states. For example, Ercikan found that the NAEP grade 8 scale score corresponding to a norm-referenced normal-curve-equivalent score of 90 ranged from a low of 305 to a high of 325 among the four states. A 20-point swing on the NAEP scale is quite large in comparison to the magnitude of differences that are seen from one administration of NAEP to the next within a state. Clearly, the links were not invariant across states as they would be expected to be if the linkings met the requirements of equating.

In a separate study, Linn and Kiplinger (1995) investigated the linking of the norm-referenced tests used in four states to the 1990 NAEP results for those same states, using equipercentile methods. They then used the 1992 NAEP results to evaluate the stability of the linkings across time. They also investigated the invariance of the linkings of the state tests with the 1990 NAEP data when the linkings were performed separately for males and females in two of the states where gender identification was possible. They found that the linkings were not invariant for males and females or across years. Rather, the linking results held only for the time and subpopulation of students used to link the two assessments and did not hold up for other subpopulations or years.

In all four states the observed 1992 NAEP scale score differed from the score that was estimated based on the 1990 linking of NAEP to the norm-referenced test. The differences were more than twice the standard error at several different percentile points in each state and were as large as 10 NAEP scale points. The changes in the linkings from 1990 to 1992 may have been due to an instructional focus on the content of the state test during that period, without a similar focus on the content of NAEP. The separate linkings for males and females also differed by more than twice the standard error at various percentile points for the two states where gender information was available for analysis. The male-female differences in linking were as large as 11 NAEP scale points.

Waltman (1997) compared two approaches to making comparisons between the Iowa Tests of Basic Skills (ITBS) and NAEP. The two assessments were linked using a social moderation approach in which the achievement level descriptions used to set performance standards for NAEP were used by judges to set performance standards (basic, proficient, and advanced) on the ITBS. An equipercentile linking of the ITBS and NAEP was also used to link the two assessments statistically. A comparison between the two linkings showed that the percentages of students scoring at the basic, proficient, and advanced levels according to the socially moderated standards were larger on the ITBS than on NAEP. As would be expected, the percentages in the achievement level categories were similar using the equipercentile linking results.

Williams, Rosa, McLeod, Thissen, and Sanford (1998) linked the grade 8 North Carolina end-of-grade (NC-EOG) mathematics test to the grade 8 NAEP mathematics assessment using a projection method. Their study was based on a special administration of a short form of the NC-EOG and two blocks of released NAEP items. Unlike the linkings performed by Ercikan (1997), Linn and Kiplinger (1995), and Waltman (1997) that relied on score distributions to perform equipercentile linkings, the Williams et al. effort used matched individual-level scores. In a second analysis, Williams et al. also used matched individual-level scores to perform a projection-based linking that predicted February 1994 NAEP results from the subsequent May 1994 NC-EOG operational administration.

Based on the results of their analyses, Williams et al. (1998) reached the following conclusions.

> The NC-NAEP linkage permits comparisons to national data and national standards. Linkage is not, however, without its problems; untestable assumptions must be made in any informative use. Chief among these problems are decisions about the use of ancillary information, and the characterization of non-sampling variation in the results (p. 294).

Despite the relatively negative results of the efforts to link state tests to NAEP just reviewed, interest in linking state assessments to NAEP has increased in the last few years. The increased interest is due, in large part, to the requirement of No Child Left Behind (NCLB) that states administer tests of mathematics and reading or English language arts to all students in grades 3 through 8 and one grade in high school each year and that they report the results of those assessments in terms of academic achievement standards. NCLB established a goal for states to have all students at the proficient level or above by 2014, but the definition of proficient performance is left to the states. Serious questions have been raised about the comparability of proficient achievement across states.

NCLB also required states to administer NAEP reading and mathematics assessments at grades 4 and 8 every other year beginning in 2003. The state NAEP results provide a common metric for comparing state achievement results. Linking state assessments to NAEP is of interest because the links would provide a means of comparing the state assessment results for different states. Thus, linking is expected to provide the basis for comparing the percent proficient results obtained by different states on their own state assessments.

As discussed above, several attempts in the 1990s to link state assessment scales to the NAEP scale using state-level data or individual data were successful for some state assessments but not for others. More recently, McLaughlin and his colleague Bandeira de Mello pioneered the linking of state assessment standards to NAEP using school-level data (McLaughlin, 1998; McLaughlin & Bandeira de Mello, 2002, 2003; McLaughlin et al., 2008). Braun and Qian (2007) have implemented a minor modification of this technology. Based on this work, the National Center for Education Statistics (NCES) released a report that linked state assessments in reading and mathematics to the 2003 and 2005 NAEP reading and mathematics assessments at grades 4 and 8 (NCES, 2007). The mapping of state proficiency standards to NAEP achievement levels in that report has received considerable attention.

Essentially, this work compares the marginal distribution of NAEP achievement scores in a state to the proportions of students reported to be meeting the state's standard on its own assessment, to identify the NAEP cut score that an equivalent proportion of students exceed. This mapping can be used to compare the proportions in a cross-classification of examinees categorized by achievement levels on NAEP and a statewide test, as shown in table 2.

Table 2. Proportions in a cross-classification of examinees categorized by achievement levels on NAEP and a statewide test.

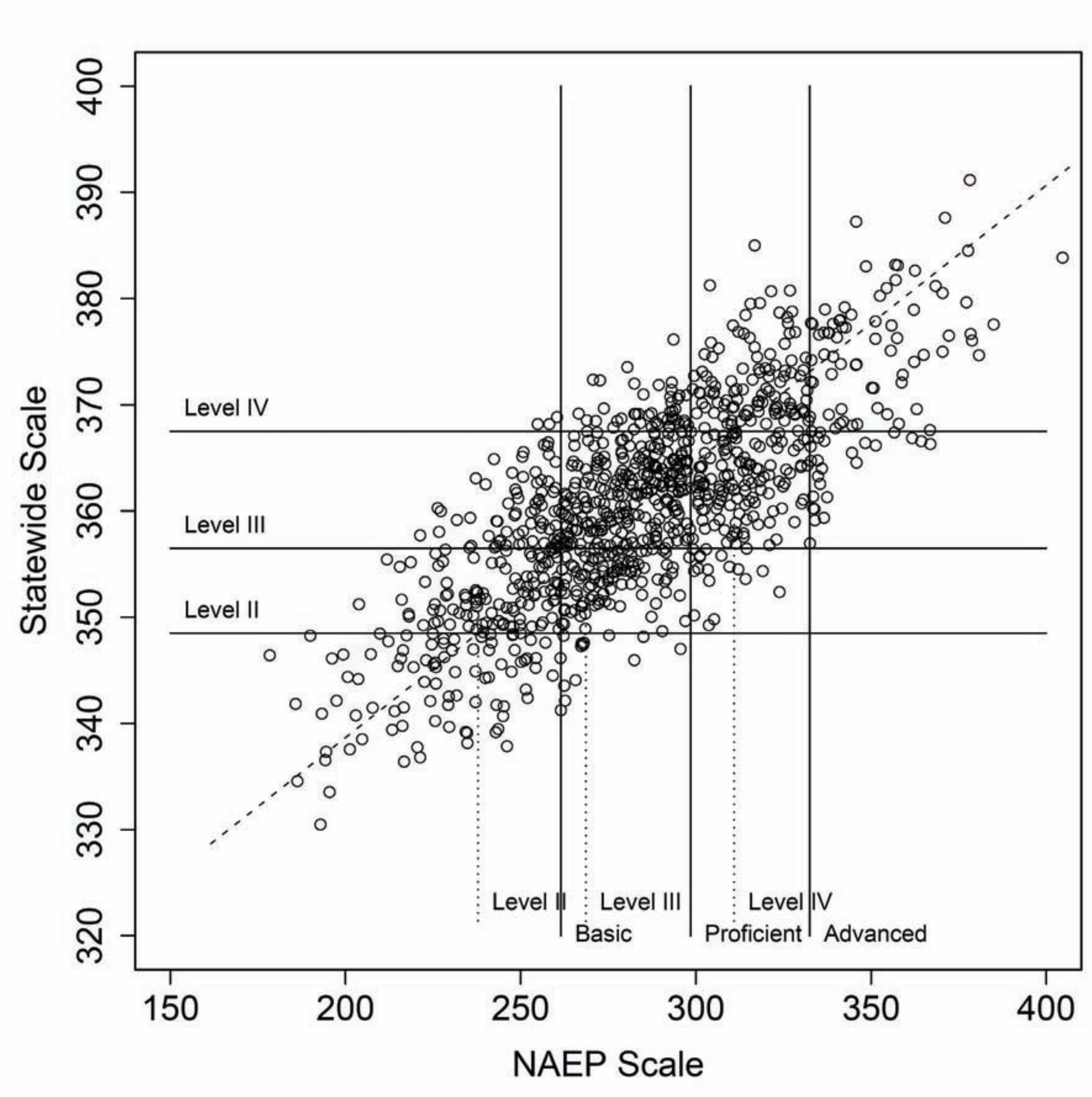
	NAEP Achievement Levels				
Statewide levels	**Below Basic**	**Basic**	**Proficient**	**Advanced**	
Level IV	*P*[IV, BB]	*P*[IV, B]	*P*[IV, P]	*P*[IV, A]	*P*[IV]
Level III	*P*[III, BB]	*P*[III, B]	*P*[III, P]	*P*[III, A]	*P*[III]
Level II	*P*[II, BB]	*P*[II, B]	*P*[II, P]	*P*[II, A]	*P*[II]
Level I	*P*[I, BB]	*P*[I, B]	*P*[I, P]	*P*[I, A]	*P*[I]
	P[BB]	*P*[B]	*P*[P]	*P*[A]	

One way to understand how this works is to consider the graphic in Figure 1, which shows the relation between scores on a hypothetical statewide test and NAEP scores. The points represent 1000 examinees;[3] the solid vertical lines represent the cut scores for the Basic, Proficient, and Advanced achievement levels on NAEP; and the horizontal lines represent the cut scores for Levels II, III, and IV on the statewide scale. The dashed line represents the principal axis, and the vertical dotted lines indicate the positions of the cut scores for Levels II, III, and IV on the NAEP scale obtained by projecting downward from the intersection of the cut scores for Levels II, III, and IV and the principal axis.

[3] To simplify the graphic for our purposes here, the points in Figure 1 represent test scores even though NAEP does not, strictly speaking, yield test scores as point estimates. Very technically, each point in Figure 1 should be shown as a (relatively narrow) horizontal distribution.

If the data were filled in, the body of table 2 would provide a numerical summary of Figure 1, showing the proportions of examinees in each of 16 cells. For example, the cell labeled P[IV, BB] would give the proportion at achievement level IV on the statewide test and Below Basic on NAEP, while the cell labeled P[III, B] would show the proportion of examinees who are at level III on the statewide test and Basic on NAEP. Note that the proportion in the first of these cells would be very small since there are only three points in Figure 1 that are in this range. By contrast, the proportion in the second cell would be relatively large.

Figure 1. Plot of the relation between scores on a hypothetical statewide test and NAEP scores



The kind of data shown in Figure 1 have only rarely been available. In order to produce them, matched individual statewide test scores and NAEP results are required. The methods used by McLaughlin and Bandeira de Mello or Braun and Qian, by contrast, do not use the individual-level data that would be required to construct Figure 1 or to populate the cells of table 2. Instead, they use the ideas described by Figure 1, through the marginal proportions shown in table 2 (which are readily available), to compute estimates of the NAEP cut-score-equivalents for statewide testing cut-scores. The estimates also depend on plausible assumptions about the form of the body of table 2, and the appearance of figure 1, but these are assumptions, not data.

In addition to his work with school-level data, McLaughlin (2001) carried out an exploration of the feasibility of constructing student-level linkages between state tests and NAEP. In 1997, state assessment scores linked to the 1996 NAEP mathematics assessment at the student level were acquired from four states with sufficient safeguards of individual student and school privacy to meet NCES standards for confidentiality. Using these data, McLaughlin projected state scores onto the NAEP scale using multiple regression and explored the structure of the error in those projections. He found correlations that were above .75 in three of the four states but lower in the fourth state. Overall, the correlations were adequate for projecting average scores for large groups of students, but involved too much error for use in estimating how individual students would have performed on NAEP. In addition, in these projections, minority status was a significant predictor. That is, the projection function was not the same for minority and non-minority students.

The four-state study formed the basis for guidelines for constructing and evaluating NAEP-state linkages. McLaughlin also generated linkages in six other states for the 1998 NAEP reading assessment, with similar results.

In 2005, as a part of the work of the NAEP Validity Studies (NVS) Panel, McLaughlin and his associates carried out student-level linkages between four state assessments and the 2003 NAEP reading and mathematics assessments in grades 4 and 8. These linkages served three separate purposes: (1) to address the question of whether using state assessment scores as a "first stage test" to assign NAEP students to item blocks closer to their achievement levels might reduce the standard errors for low performing groups; (2) to address the question of whether state assessment scores might improve the adjustment of NAEP means to take the effect of absences into account; and (3) to provide an independent verification of the levels of performance of students excluded from NAEP due to disabilities or limited English proficiency.

Student-level linkages are not always possible, and they are not needed for results that focus on higher levels of aggregation. In 1997, McLaughlin and Drori (2000) carried out a school-level linkage of state tests to NAEP in order to create a national achievement measure that could be used to investigate the correlations between various school-level measures obtained by the Schools and Staffing Survey (SASS) and school-level achievement. While SASS did not have a sufficient school sample size to support such analyses in any single state, transformation of state assessment scores onto the NAEP scale made cross-state analyses possible. These analyses yielded useful results in spite of the additional error component introduced by the linkage. For example, the study found class size to be a statistically significant correlate of achievement.

In 2004, McLaughlin and Bandeira de Mello (McLaughlin et al., 2008) constructed school-level linkages for 1998, 2000, 2002, and 2003 between state reports of percentages of students meeting standards and NAEP performance distributions in the same schools. The purpose of this linkage effort was to address the issue of whether state reports of gains and gaps were significantly different from NAEP reports of those gains and gaps. Since reports of gains and gaps in terms of percentages of students meeting a standard are a function of the placement of the standard in the distribution, the only way to validly compare state and NAEP results was to estimate the point on the NAEP scale that was equivalent to the state's standard—that is, the point at which the percentage of the NAEP distribution matched the state's report of the percentage of students meeting its standard. Thus, equipercentile linkage was used to establish the NAEP equivalents of states' standards. An important byproduct of this effort was the discovery and estimation of the great diversity in the placement of different states' standards. Results showed clearly that the percentages states reported to be meeting their standards were as much a function of the placement of their standards as the performance of their students: States with higher, or more difficult, standards (based on the linkage to NAEP) tended to report fewer students meeting those standards.

These studies give a sampling of the wide range of potential uses for databases linking NAEP and state assessment scores.

Such studies would be even more informative if data-merging arrangements could be made that allowed researchers to link individual students' statewide test scores with NAEP results. With this information, it would be possible to make plots like figure 1, and the corresponding tabulations like table 2, disaggregated by (sub)populations if that was relevant. Then we would really know how scores on the NAEP scale are related to scores on the statewide test's scale.

Results might or might not lead to a "linkage" in the sense of some score-translation system; if the relation is not invariant over (sub)populations, it probably would not. However, such validity research would answer the question "Can one describe the relations between scores obtained with one test and those obtained with the other?" This type of analysis would fall into the bottom row of table 1: answering a validity question, not creating a linkage. The relation between the scores on the two tests would not have to be linear, or homoscedastic (as it is shown in figure 1), and the proportions in table 2 might be more "spread" across the table for some achievement levels relative to others. This would all be more informative than a number-to-number translation of the cut scores from one scale to the other.

Links to International and Other National Assessments

There have been several efforts to link NAEP to other assessments over the past 15 years. Early efforts include the linking of the International Assessment of Educational Progress (IAEP) to NAEP (Pashley & Phillips, 1993; Beaton & Gonzales, 1993). Shortly after the IAEP-NAEP linkages were completed, an effort was made to link the Armed Services Vocational Aptitude Battery (ASVAB) to NAEP (Bloxom, Pashley, Nicewander, & Yan, 1995). These early linking efforts were followed by studies that linked the 1996 NAEP mathematics and science assessments to the 1995 TIMSS mathematics and science assessments (Johnson & Siengondorf, 1998), and the 2000 NAEP and the 1999 TIMSS assessments of mathematics and science (Johnson, Cohen, Chen, Jiang, & Zhang, 2005). Most recently, Phillips (2007a) published results that used a linking of NAEP and TIMSS to map the 2000 NAEP achievement levels onto the 1999 TIMSS assessment results in mathematics and science. The same linking was also used to map NAEP achievement levels onto the 2003 TIMSS assessment results.

Since all states must now participate in the NAEP mathematics and reading assessments every other year at grades 4 and 8, the linking of NAEP to an international assessment such as TIMSS automatically provides a way of comparing states as well as the nation as a whole to other countries. In a separate report, Phillips (2007b) used the results of his earlier linking of NAEP and TIMSS mathematics and science assessments to compare the 2005 and 2007 NAEP mathematics and science achievement for states to the performance of other nations on the 2003 TIMSS assessments in those subjects. According to the linking, the mean score on the TIMSS mathematics scale was above the NAEP proficient range for five countries in 2003, but no state had a mean NAEP mathematics score in the proficient range in 2007. (However, the mean for the highest-scoring state, Massachusetts, was only one point below the NAEP proficient cut score.)

Although NAEP, IAEP, and TIMSS all assess student achievement in mathematics and science and are fairly similar in content coverage, their content specifications are by no means identical. Thus, they do not satisfy one of the key requirements for equating that assessments must measure the same construct (Dorans & Holland, 2000; Holland, 2005, 2007; Linn, 1993; Mislevy, 1992). It is not easy to determine if two assessments measure the same construct, but at a minimum, the assessments should be developed from the same content specifications and use the same item types. NAEP, IAEP, and NAEP were not developed from the same content specifications. This is a major reason that the assessments should not be thought of as strictly interchangeable (as they would be if they satisfied the requirements of equating). The ASVAB differs from NAEP in terms of content coverage even more than do either IAEP or TIMSS.

What are the consequences of ignoring likely differences in the structure of two assessments and linking them anyway, and then treating the results as though they are interchangeable? A likely consequence is a lack of invariance of the linking across groups or subpopulations. Why are we concerned about this? If the two tests measure different "combinations" of constructs, or "balance" those constructs differently, it could easily be the case that one group or subpopulation performs relatively better on one (set of) constructs (and therefore on one of the tests) while another group may exhibit strengths on components of achievement that are emphasized more on the other assessment. As a result, the inferences that would be drawn through linking about performance on a test that was not given may err in one direction for one group or subpopulation, and the other direction for other subgroups. Because drawing inferences about the performance of groups or subpopulations is often an important use of an achievement test's results, those would be regrettable errors. This situation also makes overall results a function of the precise "mix" of subpopulations that happen to be in the sample—also undesirable.

A Framework to Consider Linking Utility and Validity

To consider the issues of linking utility and validity, it is essential to start by laying out the potential uses of linking. It is important, for example, not to limit possible uses of linking to those associated with high stakes test equating. As is evident from the caveats provided by the ACT and College Board, linkings such as the ACT-SAT concordance tables do not purport to provide precise estimates of the score that a student who took one test would get if they took the other test. Unlike the equating results for alternate forms of either the ACT or the SAT that are intended to yield interchangeable scores, the concordance table are intended only to provide rough approximations.

State assessments, NAEP, the measures to be used in HSLS:09, and international assessments such as TIMSS or PISA all differ from each other in ways that violate one or more of the requirements of equating. Nonetheless, there is a great interest in making comparisons across state, national, and international assessments that seem to be possible through the linkage of different assessments to each other. The question is whether or not linking through the various procedures that have been employed produces results that, while not strictly equivalent, are sufficiently comparable to make useful and valid inferences.

There are at least six different kinds of questions that can be answered through test linking that cannot be answered without linking.

1. Does one set of examinees have a reliably higher level of achievement than a second set of examinees, when the two sets of examinees take different tests?

2. Is the test or performance standard in place in one jurisdiction more difficult than the test or standard in place in a second jurisdiction, when there are no individuals who take both tests?

3. Is the test in place in one jurisdiction more sensitive to a particular variation in the population than the test in place in a second jurisdiction, when there are no individuals who take both tests?

4. Is a particular test a valid predictor of a critical outcome, when there are no individuals for whom both scores on this test and outcomes are known, but there are such data for a second test?

5. What is the cutoff score on a particular test that is equivalent to an existing cutoff on a second test?

6. What is the population estimate for achievement on a test when a non-random subpopulation is missing scores but has scores on a second test?

The utility of the answers to these questions based on linking depends on the kinds of decisions that are to be made and the increase in the positive outcomes of those decisions that can be achieved through linking, when compared to decisions based on no information. Of course, estimating a target test's scores from linkage to a second test generally involves a larger error of measurement than would be the case if scores on the target test were available, but the critical criterion for linking is the likelihood that a *conclusion* reached based on linking would be in error.

Degree of Accuracy

The degree of accuracy needed for a linking of two tests depends on the uses and interpretations to be made of the results. Some comparisons are so general in nature or so loosely defined that they can be justified without the need of any formal linking between two tests. For example, when it was reported by McCabe (2006) that 89 percent of grade 4 students scored at the proficient level or above on the Mississippi state reading assessment in 2005 while only 50 percent of the grade 4 students in Massachusetts scored at the proficient level or above on the Massachusetts state reading assessment, it is obvious to those with any knowledge of results on educational assessments in those states that the two state tests and/or the two state definitions of proficient performance differ in stringency. The same could be said for the Colorado and Missouri state assessments of grade 4 mathematics in 2005 based on the fact that 90 percent of the Colorado fourth graders were reported to be proficient or above compared to only 43 percent of fourth graders in Missouri (McCabe, 2006, p. 79).

On the other hand, a comparison of the performance of grade 8 students on the 2005 state mathematics tests where 69 percent of the students in Idaho compared to 70 percent of the students in West Virginia were at a proficient level or above (McCabe, 2006, p. 79) would not by itself justify the conclusion that the differences in percentages were the result of differences in the stringency of the tests or in the definition of proficient achievement rather than real differences in mathematics achievement or other factors. More information than the percentage of students who scored at the proficient level or above would be required to make a valid comparison.

NAEP provides one natural source of additional information. In 2005, 30 percent of the public school students in Idaho scored at the proficient level or above on the grade 8 mathematics NAEP assessment, compared to only 17 percent for West Virginia (Perie, Grigg, & Dion, 2005, p. 16). The comparison of the Idaho and West Virginia grade 8 state assessment results to the NAEP results is hardly a linking, but it illustrates the sort of comparison that is facilitated by a formal linking of state assessments to NAEP. Such a comparison, even if based on a sophisticated formal linking such as the ones conducted by McLaughlin and his colleagues (e.g., McLaughlin, Bandeira de Mello, Blankenship, Chaney, Esra, Hikawa, Rojas, William, & Wolman, 2008) or by Braun & Qian (2007), would not prove that the West Virginia test or performance standards are less stringent than those in Idaho, but they make that interpretation more plausible.

Percent Above Cut (PAC)

The *No Child Left Behind (NCLB) Act of 2001* gave new force to the practice of setting performance standards that had been encouraged by IASA and the standards movement. The outcome of primary concern under NCLB is the percentage of students who score at the proficient level or above on state assessments. Standards of basic, proficient, and advanced levels of achievement become operational when cut scores are set using one of several different standard setting techniques that involve panels of judges. The cut scores are used to produce percentage above cut (PAC) statistics for each performance standard. The PAC for students who score above the proficient cut score is critical for determining where schools meet or fail to meet adequate yearly progress targets each year.

The PAC statistics are intuitively appealing as ways to communicate student achievement and do not require the use of a scale score, which is likely to be poorly understood by non-technical audiences. Because the setting of cut scores for proficient and other performance standards on state assessments required by NCLB is left to the states, however, the meaning of proficient performance varies widely across states and between the levels set by states and the levels established for NAEP (see, for example, McLauglin et al., 2008). Equally important is the fact that PAC statistics are poorly suited for two of the purposes for monitoring state assessment results to satisfy the major goals of NCLB: (1) the tracking of trends in achievement, and (2) the tracking of changes in achievement gaps.

Holland (2002) clearly demonstrated that PAC statistics have severe limitations for assessing the magnitude of gaps in achievement, for tracking overall trends, and for evaluating changes in gaps. It can easily happen, for example, that the gap in achievement appears large based on the PAC for the basic performance standard, while it appears small based on the PAC for the proficient performance standard. Similarly, the trends can appear to be positive with the PAC for one performance standard, but flat or negative using the PAC for another standard, and gaps that appear to be closing under one scenario can appear to be increasing using PACs for different performance standards (see, for example, Ho, 2007). Furthermore, limitations of PAC statistics for tracking progress are exacerbated when PAC statistics are used to compare trends, gaps, or changes in gaps for two different assessments (e.g., a state assessment and NAEP). Effect-size statistics have better properties than PAC statistics for evaluating gaps and monitoring trends.

For many purposes the use of metric-free measures of trends and gaps suggested by Ho and Haertel (2006) provides a dependable means not only of tracking trends in overall achievement and in achievement gaps, but of comparing trends for different assessments (see also Ho, 2007). The probability-probability (PP) plots advocated by Ho and Haertel (2006) are curves in the unit square that show the proportions from each of two distributions that are below a given score, x. The PP plots remain unaffected by monotone transformations of the test scale score. Identical cumulative frequency distributions would yield a PP curve that is a 45 degree diagonal from 0,0 to 1,1 in the unit square. PP curves that are all above or all below the diagonal result when the distributions are stochastically ordered (i.e., when the cumulative frequency of one distribution is always greater than or equal to the other). Cumulative frequency distributions that are not stochastically ordered result in PP curves that cross the diagonal.

A useful, metric-free statistic, which Ho and Haertel (2006) refer to as V (for deViation), is the area between the PP curve and the diagonal in a PP plot. V is equal to the probability that a randomly drawn observation from the distribution represented by the vertical axis of the PP plot is above a randomly drawn observation represented by the horizontal axis (Ho, 2007). As Ho (2007) has illustrated, V statistics can be useful in comparing trends from different assessments such as a state assessment and NAEP.

Threats to Utility and Validity of Linking

Several differences between assessments can threaten the utility and validity of a linking. Other things being equal, the trustworthiness of a linking will be greater for assessments that are aligned to similar content standards and assessment frameworks than for assessments that are aligned with dissimilar content standards and assessment frameworks (Feuer et al., 1999; Ho & Haertel, 2007). Major differences in content standards and assessment frameworks can result in linkings that have little utility or validity. Major differences in the format of assessment items can also reduce the utility and validity of a linking. For example, the linking of an assessment that uses only multiple-choice items to one that relies heavily on constructed response items is suspect (Feuer et al., 1999).

For assessments that are intended to measure a student's maximum performance to yield valid interpretations, students must be motivated to do their best. Clearly, if students taking one assessment are highly motivated, while students taking another assessment are not motivated, then a linking of the two assessments will be less trustworthy than it would be if the motivational conditions were more comparable (Bloxom et al., 1995; Feuer et al., 1999). "If the data for two assessments in a linkage study arise from circumstances under which the examinees are more motivated on one than the other, the average level of the projection might be too high or too low relative to that which would be obtained if the second test had actually been administered" (Thissen, 2007, pp. 306-307). The potential effect of motivation is one of the major reasons why educational measurement experts generally prefer to try out new items using field tests where the new items are embedded in an operational form of the test rather than relying on stand-alone field tests.

Even when a linking is conducted in situations that result in interpretations that are useful and have a high degree of validity at one point in time, the linking may not yield valid interpretations about the scores at a different point in time: "Inferences based on … links about years other than the one in which the link is calculated are not warranted because of likely failure of the invariance of the linking functions across time" (Koretz, 2007, p. 348).

An Alternative to Linking

If questions about the relations between the NAEP scale and the scales of other tests[4] were phrased as validity research ("What is the relation of results obtained from <the other test> with NAEP results?") as opposed to an *a priori* specification that <the other test> should be *linked* with the NAEP scale, more useful results could be obtained and controversy could be avoided. Controversy arises when untestable assumptions are made during linking and no data are available to resolve questions about those assumptions. Data collection designed to explicate the relation between NAEP results and those obtained with <the other test> would answer questions about the form of the relationship between the two score scales, and the invariance (or lack of it) in that relation across subgroups. Result for both tests would need to be collected either at the individual student level or for aggregates such as schools. It would then be possible to evaluate the relationship between the two sets of scores in a manner similar to that which is displayed in Figure 1. Instead of a linkage being said to "fail" due to a lack of invariance of the relation across subpopulations, differences between

[4] Other tests may be statewide or international assessments, or research instruments.

subpopulations in the relation between the tests would be part of the results of the validity research.

Such research would necessarily involve designs in which both tests were administered to the same units of analysis (most straightforwardly, students; however, units that are higher levels of aggregation may serve some purposes). The sample sizes would need to be sufficiently large, and the design planned, so that the invariance (or lack of invariance) in the relation between the two tests could be examined over some potentially relevant subgroups. These requirements mean that such research could rarely, if ever, be carried out using only the publicly reported results from large-scale assessments. Special data collection would be required.[5]

The rewards for such research would be substantial, as it would simultaneously explicate the nomothetic span of both NAEP and <the other test>, and, to the extent possible, provide answers to some of the questions listed on page 17. For example, a version of table 2 completely filled with empirical data, or better yet, an empirically filled figure 1, would make clear the relation between cut scores and difficulties of two tests (questions 2 and 5); (sub)population differences in those displays would answer question 3. Given those answers, further answers to question 1, comparing achievement levels for groups, may be possible. An answer to question 6 would require a specialized version of these analyses with at least some representation of the missing data. Question 4 is the classic question of validity, which should be answered for any test.

We recommend that this “new perspective” replace the search for a grail that linkage across dissimilar tests has become.

[5] One such possibility concerns the first coincidence of TIMSS and PIRLS (which are administered together only once every 20 years) with the collection of NAEP mathematics and reading at Grades 4 and 8 in 2011. This coincidence provides an unprecedented opportunity to gather data on mathematics, science, and reading achievement data on the same individuals in many countries, and—potentially—to utilize the NAEP link to expand our understanding of U.S. achievement in a global context.

References

Angoff, W. H. (1964). Technical problems of obtaining equivalent scores on tests. *Journal of Educational Measurement*, *1*, 11-13.

Beaton, A. E. & Gonzales, E. J. (1993). Comparing the NAEP trial state assessment results with the IAEP international results. In L. A. Shepard, R. Glaser, R. Linn, & G. Bohrnstedt (Eds). *Setting performance standards for student achievement: Background studies*, Stanford, CA: The National Academy of Education.

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.

Braun, H. I. and Qian, J. (2007). An enhanced method for mapping state standards onto the NAEP scale. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 313-338). New York: Springer.

Dorans, N. J. (2004). Equating, concordance, and expectation. *Applied Psychological Measurement*, *28*, 227-246.

Dorans, N. J. & Holland, P. W. (2000). Population invariance and the equitability of tests: Basic theory and the linear case. *Journal of Educational Measurement*, *37*, 281-306.

Dorans, N. J., Lyu, C. F., Pommerich, M., & Houston, W. M. (1997). Concordance between the ACT assessment and the recentered SAT I sum scores. *College and University, 3*(2), 24-34.

Embretson (Whitely), S. (1983). Construct validity: construct representation vs. nomothetic span. *Psychological Bulletin*, *93*, 179-197.

Ercikan, K. (1997). Linking statewide tests to the National Assessment of Educational Progress: Accuracy of combining test results across state. *Applied Measurement in Education*, *10*, 145-159.

Feuer, M. J. (2005). E Plurbus 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., 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.

Hambleton, R. K., Jaeger, R. M, Koretz, D., Linn, R. L., Millman, J., & Phillips, S. E. (1995). *Review of the measurement quality of the Kentucky Instructional Results Information System, 1991-*

1994. Frankfort, KY: Office of Education Accountability, Kentucky General Assembly.

Ho, A. D. (2007). Discrepancies between score trends from NAEP and state tests: A scale invariant perspective. *Educational Measurement: Issues and Practice*, *26*(4), 11-20.

Ho, A. & Haertel, E. H. (2006). *Metric-free measures of test score trends and gaps with policy-relevant examples*. Technical Report #665. University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Los Angeles, CA.

Ho, A. & Haertel, E, (2007). *(Over)-interpreting mappings from state performance standards onto the NAEP scale*. Retrieved from the Council of Chief State School Officers: www.ccsso.org/content/PDFs/Ho%20Brief%20final.pdf.

Holland, P. W. (2002). Two measures of change in the gaps between CDFs of test score distributions. *Journal of Educational and Behavioral Statistics*, *27*(1), 3-17.

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.

Holland, P. W. (2007). Framework and history for score linking. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 5-30). New York: Springer.

Holland, P. W. & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 187-220). Westport, CT: Praeger Publishers.

Johnson, E. G., Cohen, J., Chen, W.-H., Jiang, T., & Zhang, Y. (2005). *2000 NAEP-1999 TIMSS linking report* (Publication No. 2005-01). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Johnson, E. G., Siengondorf, A. (1998). *Linking the National Assessment of Educational Progress and the Third International Mathematics and Science Study: Eighth grade results* (Publication No. NCES 98-500). Washington, DC: National Center for Education Statistics.

Kolen, M. J. (2004). Linking assessments: Concept and history. *Applied Psychological Measurement*, *28*, 219-226.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking. Methods and practice* (2nd ed.). New York, Springer.

Koretz, D. (2007). Using aggregate-level linkages for estimation and validation: Comments on Thissen and Braun & Qian. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 339-353). New York: Springer.

Koretz, D. & Barron, S. I. (1998). *The validity of gains in scores on the Kentucky Instructional Results Information System (KIRIS).* (MR-1014-EDU). Santa Monica, CA: RAND.

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 tests to the National Assessment of Educational Progress: Stability of results. *Applied Measurement in Education*, *8*, 135-155.

McCabe, M. (2006). The state of the states. *Education Week, Quality Counts: A decade of standards based education, A+ 10*, *25*(17), January 5, pp. 72-96.

McLaughlin, D. H. (2001) *Study of the linkages of 1996 NAEP and state mathematics assessments in four states.* Washington, DC: National Center for Education Statistics. (NCES 2001-481).

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.

McLaughlin, D., & Bandeira de Mello, V. (2003). *Comparing State Reading and Math Performance Standards Using NAEP*. Paper presented at the annual National Conference on Large-Scale Assessment, San Antonio.

McLaughlin, D. H. & Drori, G. (2000) *School-level correlates of academic achievement: Student assessment scores in SASS public schools.* Washington, DC: U.S. Department of Education, (NCES 2000–303).

McLaughlin, D., Bandeira de Mello, V., Blankenship, C., Chaney, K., Esra, P., Hikawa, H., et al. (2008). *Comparison between NAEP and state mathematics assessment results 2003, volumes 1 and 2. Research and development report.* Washington, DC: National Center for Education Statistics.

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

National Center for Education Statistics (2007). *Mapping 2005 state proficiency standards to NAEP scales.* (NCES 2007-482). U.S. Department of Education, Washington, DC.

National Governors Association, the Council of Chief State School Officers, and Achieve, Inc. (2008). *Benchmarking for success: Ensuring U.S. students receive a world-class education.* Authors: Washington, DC.

Perie, M., Grigg, W., & Dion, G. (2005). *The nation's report card: Mathematics 2005.* (NCES 2006-453). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Phillips, G. W. (2007a). *Expressing international educational achievement in terms of U.S. performance standards: Linking NAEP achievement levels to TIMSS*. Washington, DC: American Institutes for Research.

Phillips, G. W. (2007b). *Chance favors the prepared mind: Mathematics and science indicators for comparing states and nations.* Washington, DC: American Institutes for Research.

Pommerich, M., Hanson, B. A., Harris, D. J., & Sconing, J. A. (2004). Issues in conducting linkages between distinct tests. *Applied Psychological Measurement*, *28*, 227-246.

Sawyer, R. (2007). Some further thoughts on concordance. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 217-230). New York, NY: Springer.

Thissen, D. (2007). Linking assessments based on aggregate reporting: Background and issues. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 287-312). New York, NY: Springer.

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., Rosa, K. R., McLeod. L. D. Thissen, D., & Sanford, E. E. (1998). Projecting to the NAEP scale : Results from the North Carolina end-of-grade testing program. *Journal of Educational Measurement*, *35*, 277-296.