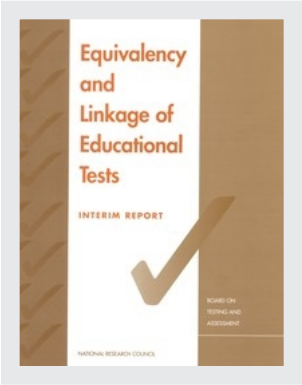


# Equivalency and Linkage of Educational Tests: Interim Report

**DETAILS**

46 pages | 8.5 x 11 | PAPERBACK
ISBN 978-0-309-06177-3 | DOI 10.17226/6209



**AUTHORS**

Michael J. Feuer, Paul W. Holland, Meryl W. Bertenthal, F. Cadelle Hemphill, and Bert F. Green, Editors; Committee on Equivalency and Linkage of Educational Tests, National Research Council

# Equivalency and Linkage of Educational Tests

## INTERIM REPORT

Michael J. Feuer, Paul W. Holland,
Meryl W. Bertenthal, F. Cadelle Hemphill, and Bert F. Green, Editors

Committee on Equivalency and Linkage of Educational Tests



Board on Testing and Assessment

Commission on Behavioral and Social Sciences and Education

National Research Council

NATIONAL ACADEMY PRESS
Washington, D.C. 1998

NOTICE: The project that is the subject of this report was approved by the Governing Board of the National Research Council, whose members are drawn from the councils of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine. The members of the committee responsible for the report were chosen for their special competences and with regard for appropriate balance.

The National Academy of Sciences is a private, nonprofit, self-perpetuating society of distinguished scholars engaged in scientific and engineering research, dedicated to the furtherance of science and technology and to their use for the general welfare. Upon the authority of the charter granted to it by the Congress in 1863, the Academy has a mandate that requires it to advise the federal government on scientific and technical matters. Dr. Bruce M. Alberts is president of the National Academy of Sciences.

The National Academy of Engineering was established in 1964, under the charter of the National Academy of Sciences, as a parallel organization of outstanding engineers. It is autonomous in its administration and in the selection of its members, sharing with the National Academy of Sciences the responsibility for advising the federal government. The National Academy of Engineering also sponsors engineering programs aimed at meeting national needs, encourages education and research, and recognizes the superior achievements of engineers. Dr. William A. Wulf is president of the National Academy of Engineering.

The Institute of Medicine was established in 1970 by the National Academy of Sciences to secure the services of eminent members of appropriate professions in the examination of policy matters pertaining to the health of the public. The Institute acts under the responsibility given to the National Academy of Sciences by its congressional charter to be an adviser to the federal government and, upon its own initiative, to identify issues of medical care, research, and education. Dr. Kenneth I. Shine is president of the Institute of Medicine.

The National Research Council was organized by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of furthering knowledge and advising the federal government. Functioning in accordance with general policies determined by the Academy, the Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in providing services to the government, the public, and the scientific and engineering communities. The Council is administered jointly by both Academies and the Institute of Medicine. Dr. Bruce M. Alberts and Dr. William A. Wulf are chairman and vice chairman, respectively, of the National Research Council.

The study was supported by Grant No. ED-98-CO-0005 between the National Academy of Sciences and the U.S. Department of Education. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the view of the organizations or agencies that provided support for this project.

International Standard Book Number 0-309-06177-6

Additional copies of this report are available from:
National Academy Press
2101 Constitution Avenue N.W.
Washington, DC 20418
Call 800-624-6242 or 202-334-3313 (in the Washington Metropolitan Area).

This report is also available on line at **http://www.nap.edu**

Printed in the United States of America

## COMMITTEE ON EQUIVALENCY AND LINKAGE OF EDUCATIONAL TESTS

PAUL W. HOLLAND (*Chair*), School of Education, University of California, Berkeley
ROBERT C. CALFEE, School of Education, Stanford University
JOHN T. GUTHRIE, Human Development, University of Maryland, College Park
RICHARD M. JAEGER, School of Education, University of North Carolina, Greensboro
PATRICIA ANN KENNEY, Learning Research and Development Center, University of Pittsburgh
VONDA L. KIPLINGER, Colorado Department of Education
DANIEL M. KORETZ, RAND, Washington, D.C.
FREDERICK C. MOSTELLER, Department of Statistics, Harvard University
PETER J. PASHLEY, Law School Admission Council, Newtown, Pennsylvania
DORIS REDFIELD, Educational Consultant, Richmond, Virginia
WILLIAM F. TATE, Department of Curriculum and Instruction, University of Wisconsin, Madison
DAVID THISSEN, Department of Psychology, University of North Carolina, Chapel Hill
EWART A.C. THOMAS, Department of Psychology, Stanford University
LAURESS L. WISE, Human Resources Research Organization, Alexandria, Virginia

ROBERT L. LINN, *ex officio*, Board on Testing and Assessment, National Research Council; School of Education, University of Colorado

MICHAEL J. FEUER, *Study Director*
BERT F. GREEN, *Senior Technical Advisor*
MERYL W. BERTENTHAL, *Senior Research Associate*
F. CADELLE HEMPHILL, *Senior Research Associate*
LISA D. ALSTON, *Senior Project Assistant*

## BOARD ON TESTING AND ASSESSMENT

# Acknowledgments

This report culminates the first six months of a nine-month effort that would not have been imaginable without the extraordinary contributions of many individuals.

Above all, we are extremely grateful to the members of the committee, who understood both the urgency and significance of their charge, gave generously of their expertise and time, and met the highest standards of the 130-year-old tradition of the National Academy complex in providing voluntary scientific advice to the government through the National Research Council (NRC). As chair of the Board on Testing and Assessment (BOTA), Robert Linn again provided impeccable judgment, wise counsel, and virtually unlimited time. We thank Bob and other members of BOTA for their superb stewardship of this (and other) BOTA projects.

At the staff level, we express special thanks to Meryl Bertenthal and Cadelle Hemphill, relative newcomers to the NRC, who quickly, gracefully, and effectively mastered the many aspects of their new jobs; to Lisa Alston, for her absolutely superb and unflappable administrative support; to Bert Green, whose wisdom born of experience and scholarship is reflected throughout the report; and to Nancy Kober, for her fine editorial and substantive judgment. Other BOTA staff—Bob Rothman, Karen Mitchell, Patricia Morison, Viola Horek, Naomi Chudowsky, Lee Jones, Kim Saldin, Alix Beatty, Allison Black, and Steve Baldwin—offered advice and support at various stages of writing and rewriting, and came through again as an invaluable team.

Barbara Torrey, executive director of the Commission on Behavioral and Social Sciences and Education (CBASSE), and Sandy Wigdor, director of the CBASSE's Division on Education, Labor, and Human Performance, have been sources of great encouragement in this fast-track study and paved many paths from committee formation through report review. Finally, extra special thanks to Eugenia Grohman, the CBASSE associate director of reports, our guardian angel.

This report has been reviewed by individuals chosen for their diverse perspectives and technical expertise, in accordance with procedures approved by the NRC's Report Review Committee. The purpose of this independent review is to provide candid and critical comments that will assist NRC in making the published report as sound as possible and to ensure that the report meets institutional

standards for objectivity, evidence, and responsiveness to the study charge. The content of the review comments and draft manuscript remain confidential to protect the integrity of the deliberative process.

We thank the following individuals, who are neither officials nor employees of the NRC, for their participation in the review of this report: Bruce Bloxom, Naval Postgraduate School (retired), Monterey, California; Robert Brennan, College of Education, University of Iowa; Arthur S. Goldberger, Department of Economics, University of Wisconsin; Lyle V. Jones, L.L. Thurstone Psychometric Laboratory, University of North Carolina, Chapel Hill; Lincoln E. Moses, Department of Statistics (emeritus), Stanford University; Stephen P. Raudenbush, School of Education, University of Michigan; Henry W. Riecken, Department of Behavioral Sciences (emeritus), University of Pennsylvania Medical Center; Richard Shavelson, School of Education, Stanford University; Mark Wilson, School of Education, University of California, Berkeley.

While the individuals listed above provided many constructive comments and suggestions, responsibility for the final content of this report rests solely with the authoring committee and the NRC.

Paul W. Holland, *Chair*
Michael J. Feuer, *Study Director*

# Contents

# Executive Summary

In November 1997 Congress requested that the National Research Council study the feasibility of developing a scale to compare ("link") scores from existing commercial and state tests to each other and to the National Assessment of Educational Progress (NAEP) (see P.L. 105-78). The Committee on Equivalency and Linkage of Educational Tests is carrying out the requested study. This report presents the committee's general findings and conclusions; it will be followed by the committee's more detailed final report in fall 1998.

The request for the study arose in the context of congressional debate about the proposed voluntary national tests. The committee assumes that the development of a method to link and compare the full array of existing tests is seen as a possible substitute for the development of new national tests. Although our findings and conclusions may be relevant to certain technical issues in the debate over the national tests, we take no position on their overall technical or policy merits.

We have reviewed empirical evidence of two basic types: on the diversity in content, usage, and purposes of educational testing in the United States and on statistical and other technical aspects of creating valid linkages among various types of educational tests.

Our findings from this review raise fundamental questions about the feasibility of linkage as envisioned by the committee's charge. Although it may be technically possible to establish links between tests that are highly similar in design, format, content emphasis, difficulty level, and intended use, those conditions do not apply to the increasingly diverse and complex array of state and commercial tests in the nation's 50 states and more than 15,000 school districts. In addition, those tests differ substantially from NAEP, which is designed specifically to provide scores for groups of students and not for individual students and which differs in other fundamental ways from almost all state and local tests.

Thus, we have reached two principal conclusions:

**Comparing the full array of currently administered commercial and state achievement tests to one another, through the development of a single equivalency or linking scale, is not feasible.**

**Reporting individual student scores from the full array of state and commercial achievement**

**tests on the NAEP scale, and transforming individual scores on these various tests and assessment into the NAEP achievement levels is not feasible.**

This interim report synthesizes the committee's review of empirical evidence that supports our conclusions. The final report will provide additional detail on the diverse landscape of educational testing in the United States, the special problems of establishing valid linkages with NAEP, and the implications of prior and ongoing research on technical aspects of linkage and equivalency. The final report will also consider the possibility of establishing criteria with which to evaluate the feasibility of linking subsets of the diverse and increasing number of tests currently used in the nation's schools.

# Introduction

In Public Law 105-78, enacted November 13, 1997, the U.S. Congress called on the National Research Council to "conduct a feasibility study to determine if an equivalency scale can be developed that would allow test scores from commercially available standardized tests and state assessments to be compared with each other and the National Assessment of Educational Progress" (NAEP) (Sec. 306). Simply stated, the question before the committee is whether reliable relationships ("linkages") can be established between the scores obtained from existing commercially produced tests, state educational assessments, and NAEP.

To carry out this charge, the National Research Council, through its Board on Testing and Assessment, established the Committee on Equivalency and Linkage of Educational Tests in January 1998. This interim report, which is required by the legislation, presents the committee's findings and conclusions on the basis of its work to date.

To accomplish its work, the committee is analyzing data on state educational assessments, consulting with other experts and practitioners involved in test design and implementation, and reviewing past and current research on linking tests. As stipulated by the law authorizing the study, we are conferring with representatives of the House and Senate education committees, the White House, the National Assessment Governing Board, the National Governors' Association, and the National Conference of State Legislatures.

The request for the study arose in the context of congressional debate about the proposed voluntary national tests. Although our findings and conclusions may be relevant to certain technical issues in the debate over the national tests, we take no position on their overall technical or policy merits.

Under its charge, the committee will continue to analyze information and data and deliberate for 3 more months. Our work will conclude with a final report in September 1998. That report will present a more detailed research review, explain more fully the issues involved in test linking, and elaborate the findings and conclusions in this interim report.

# Purpose and Scope of the Study

## POLICY CONTEXT

Why does it matter to anyone other than testing experts whether the results of state and commercial tests can be linked to a common scale? Although test linkage is a highly technical issue, the question posed by Congress reflects a broad, underlying goal held by many Americans to know more about how individual students in the United States are performing in relation to high national or international benchmarks of performance. Many people believe that students and teachers would also profit from knowing how a student's performance in key content areas compares with the performance of other students, other schools, other states, and other countries (see, e.g., Rose, Gallup, and Elam, 1997). Many people believe that if linkages among different tests enabled such comparisons they would help to spur improvements in schooling at the state and local levels (see, e.g., Achieve, 1998). Others hold different views of the utility of this type of information (see, e.g., Jones, 1997).

Existing assessments of student performance are diverse and are guided by different purposes (see, e.g., Bond, 1995). The committee recognizes in the legislation that requested this study a desire to bring about greater comparability among tests in the United States, while upholding traditions of state and local control of education and respecting the substantial public and private investments that have been made in developing educational tests and assessments. In a word, we interpret the charge to us as an expression of Congress's desire to forge some unity of interpretation within a heterogeneous system of testing and assessment. The focus of this report is whether an equivalency scale can be developed to accomplish that goal, the educational assessment equivalent of "e pluribus unum."

## PURPOSES OF LINKAGE

In formulating the question for this study, Congress was not explicit about the purposes of the proposed linkages. However, because the study originated in the vigorous debate about the President's proposal for voluntary national testing, the committee assumes that Congress sees linkage as a possible

substitute for the voluntary national tests. For example, if linkages could be developed that would permit the scores of individual students on existing tests and assessments to be compared with each other and reported in terms of the achievement levels used by NAEP, it might show which state tests are less challenging than others and whether individual children are reaching achievement levels defined by NAEP. Moreover, the information might be used by parents to work for improvements in their schools and school districts (Smith, Stevenson, and Li, 1998).

## SCOPE OF THIS STUDY

The primary focus of the committee's examination is linkage among the tests and assessments currently used by states and districts to measure individual students' educational performance. The committee uses the terms "test" and "assessment" interchangeably, following Shepard (1994). If there is a difference between these two terms it is one of emphasis: a test usually refers to a particular coherent testing instrument; an assessment is more likely to refer to a system that involves more than one test.

This interpretation of our charge has led the committee to focus its deliberations on two principal issues:

1. the degree to which existing state and commercial tests and assessments can be linked to each other on a common scale, permitting individual scores from different tests and assessments to be compared; and
2. whether scores on existing state and commercial tests and assessments can be interpreted in terms of NAEP's achievement levels, so that parents can know how well their children are doing as measured against national benchmarks.

The committee has a large field of inquiry and a short time frame in which to analyze evidence and arrive at conclusions. The committee is examining a substantial amount of data about selected state and commercial tests and assessments that are likely candidates for the types of linkage suggested by the legislation. We are specifically investigating:

- common uses of these tests;
- diversity in their content and format;
- their measurement properties, such as their difficulty and the reliability of their scores (by reliability we refer to their consistency over time);
- the degree to which state and district tests change over time; and
- the degree to which state policies affect uses and interpretations of test results.

To evaluate the methodological and technical issues involved in establishing linkages between these kinds of tests, the committee is reviewing past efforts to link different tests and assessments to each other and to NAEP. We are giving particular attention to the purposes of these linkages as we explore research on alternative methodologies and the validity of inferences drawn from the linked results. Where possible, we illustrate our empirical findings with examples from math and reading tests, the focus of much of today's policy debate.

# Coming to Terms: Assumptions, Definitions, and Goals of Linkage

## LINKING AND EQUATING

Consistent with the scientific literature on psychological and educational measurement, the committee interprets the legislation's reference to "equivalency scale" to mean the result of equating or linking the results of different tests and assessments (see, e.g., Holland and Rubin, 1982). Throughout this report we use the term "linkage" to mean various well-established statistical methods for connecting scores on different tests and assessments to each other and for reporting them on a common scale. The goal of these methods is to enable the performance of one student on one test to be compared with performances of other students on different tests (Mislevy, 1992; Linn, 1993). Box 1 describes these linkage methods.

Whatever the method used, there are two main technical concerns with linkages: accuracy and consistency. Accuracy is analogous to the "margin for error" in opinion polling and depends on the amount of data used in the calculations. The more test takers in a study, the higher the accuracy of the linkage. Consistency refers to the consistency of the linkage found across all of the relevant subpopulations of test takers, which depends (among other things) on the details of the tests themselves and on their relationship to the educational experiences of the test takers. In the dynamic situation of educational reform today, the relevant subpopulations may even include those students to be tested in the next few years for whom there are no data currently available in a linking study done today. The consequence of inaccuracy is a decrease in the reliability of the scores when they are placed on the scale of the test not taken by the test taker. The consequence of inconsistency is a degree of "bias," which can create disadvantages for some students and advantages for others.

Because the technical aspects of testing are unfamiliar to many readers of this report, analogies with measuring temperature may be useful. For example, test results are often reported on "scales" that are arbitrary in the same way that for temperature on the Fahrenheit scale, 32 degrees is freezing and 212 degrees is boiling—that is the 200 to 800 scale of the Scholastic Achievement Test or the 0 to 500 scale of NAEP. The analogy of the link between different scales for measuring temperature, for example,

### BOX 1. Linking Methods

**Equating** The strongest kind of linking, and the one with the most technical support, is equating. Equating is most frequently used when comparing the results of different forms of a single test that have been designed to be parallel. The College Board equates different forms of the Scholastic Assessment Test (SAT) and treats the results as interchangeable. Equating is possible if test content, format, purpose, administration, item difficulty, and populations are equivalent.

In linear equating, the mean and standard deviation of one test is adjusted so that it is the same as the mean and standard deviation of another. Equipercentile equating adjusts the entire score distribution of one test to the entire score distribution of the other. In this case, scores at the same percentile on two different test forms are equivalent. Thus, if a score of 122 on one test, X, is at the 75th percentile and a score of 257 on another test, Y, is also at the 75th percentile for the same population of test takers, then 122 and 257 are linked by the equipercentile method. This means that 75 percent of the test takers in this population would score 122 or less on test X or would score 257 or less on test Y. The linked scores, 122 and 257, have the same meaning in this very specific and widely used sense, and we would place the X score of 122 onto the scale of test Y by using the value of 257 for it. By following this procedure for each percentile value from 1 to 99, tests X and Y are linked.

Equipercentile equating is not the only method used to link tests, but it is a basic one and is closely connected to the other methods (Holland and Rubin, 1982; Peterson et al., 1989; Kolen and Brennan, 1995).

Two tests can also be equated using a third test as an anchor. This "anchor test" should have similar content to the original tests, although it is typically shorter than the two original tests. Often the anchor test is a separately timed section of the original tests. Sometimes, however, the items on the anchor test are interspersed with the items on the main tests. A separate score is computed for the responses to those items as if they were a separate test. An assumption of the equipercentile equating methodology is that the linking function found in this manner is consistent across the various populations that could be chosen for the equating. For example, the same linking function should be obtained if the population is restricted only to boys or only to girls. However, the research literature shows that this consistency is to be expected only when the tests being linked are very similar in a variety of ways that are discussed in the body of this report.

**Calibration** Tests or assessments that are constructed for different purpose, using different content frameworks or test specifications, will almost always violate the conditions required for equating. When scores from two

Fahrenheit and Celsius, is also a useful starting point for understanding what it means for test scores or scales to be linked; see Box 2.

## FEASIBILITY

The committee interprets feasibility to encompass both practicality and validity. By *practicality* we mean not only whether a linkage can be created, in the mathematical or statistical sense, but also whether the costs of doing so—the financial cost and logistical burden of collecting and analyzing the data necessary for the linkages are reasonable and manageable.

Due to the short time frame of this study, a comprehensive cost analysis could not be conducted. However, the committee has reviewed the Anchor Test Study (Loret et al., 1972), which developed an equivalency scale for eight reading subtests, a number representing almost 90 percent of reading tests used at that time, at the cost of more than $1 million. There is greater diversity and less stability among testing programs today, leading us to assume that the costs of linkage would be substantially higher. The committee's final report will address this issue in more detail; the issue will need more concen-

different tests are put on the same scale, the results are said to be comparable, or calibrated. Most of the statistical methods used in equating can be used in calibration, but it is not expected that the results will be consistent across different populations.

Two types of empirical data support equating and calibration of scores between two tests. In one type, the two tests are given to a single group of test takers. When the same group takes both tests, the intercorrelation of the tests provides some empirical evidence of equivalent content. In a second design, two tests are given to equivalent groups of test takers. Equivalent groups are often formed by giving both tests at the same time to a large group, with some of the examinees taking one test and some the other. When the tests are given at different times to different groups of test takers, equivalence is harder to assert.

Two tests can be equated or calibrated using a third test as an anchor. This method requires that one group of students takes tests A and C, while another group takes tests B and C. Tests A and B are then calibrated through the anchor test, C. For this method to be valid, the anchor test should have the same content as the original tests, although it is typically shorter than the other tests.

One relatively new equating procedure, used extensively in NAEP and many other large testing programs, depends on being able to calibrate the individual items that make up a test, rather than the test itself. Each of a large number of items about a given subject is related or calibrated to a scale measuring that subject, using a statistical theory called item response theory (IRT). The method works only when the items are all assessing the same material, and requires that a large number of items be administered to a large representative set of test takers. Once all items are calibrated, a test can be formed from a subset of the items, and be assured of being equated automatically to another test formed from a selection of different items.

**Projection** A special unidirectional form of linking can be used to predict or "project" scores on one test from scores on another test without any expectation that exactly the same things are being measured. Usually, both tests are given to a sample of students and then statistical regression methods are applied. It is important to note that projecting Test A onto Test B gives different results from projecting Test B onto Test A.

**Moderation** Moderation is the weakest form of linking. It is used when the tests have different blueprints and are given to different, nonequivalent groups of examinees. Procedures that match distributions using scores are called statistical moderation links, while others that match distributions using subjective judgments are referred to as social moderation links. In either case, the resulting links are only valid for making some very general comparisons (Mislevy, 1992; Linn, 1993).

trated attention if policy makers decide to proceed with a research, development, and implementation program to link existing tests.

Statisticians and other measurement experts link results of different tests through various methods, which include evaluating the content and character of tests being linked, collecting large amounts of data on student performance on each test, and carrying out statistical computations to identify accurate and consistent relationships between these test results (see, e.g., Linn, 1993; Mislevy, 1992). These methods will add to the expected cost of linking.

By *validity* we mean whether any linkage that is created can support the meaning and interpretation—the inferences—that users are likely to draw (see, e.g., Messick, 1989). The validity of an inference based on a linkage hinges on several considerations, including the similarity of content of the linked scales. If two tests both measure the degree to which a student has mastered a particular subject (domain of content), then the link permits comparisons of the results from both tests in similar terms. The validity of a link is supported by a strong statistical correlation of the scores from the two tests and by the consistency of linkage across population groups, such as boys and girls, African Americans and whites, or residents of New York and California.

**BOX 2. Fahrenheit, Celsius, and Educational Tests**

There is a well-known formula for linking Fahrenheit and Celsius temperatures: $F° = (9/5)C + 32°$. Thus, if one reads that Paris is suffering from a 35-degree heat wave—which may not seem very hot—one needs to multiply 35 by 9 and divide that result by 5 to get 63 and then add 32 to get a very recognizably hot 95, in degrees Fahrenheit. This formula is an example of a linking function and is analogous to what is meant by linking two test score scales. Just as one placed the Celsius value of 35 on the Fahrenheit scale and got 95 (which may be more meaningful to some people), linking can allow one to place the scores from one test on the scale of another and interpret that score or to compare it to those of test takers who took the other test. Other uses of linking assessments are to estimate how schools or districts would have performed had their students taken an assessment, such as NAEP, that they did not take.

Although the temperature measurement analogy is useful for understanding some aspects of tests, it is only a partial analogy because temperature measurement is very simple compared with the assessment of complex cognitive activities, such as reading or mathematics.

There are many possible purposes for establishing linkages among tests. The committee recognizes that the validity of a linkage varies depending on its purpose and use and on whether it is being used as intended. Linkages that provide valid and useful information for some purposes may nonetheless be inadequate, and so invalid, for others. For example, a link that is sufficiently accurate to categorize the educational quality of schools or school districts may not be sufficiently accurate and stable to classify the proficiency of individual students (Williams et al., 1995).

## USES OF LINKAGE

The committee realizes that with careful planning it is possible to establish valid links between tests that meet certain conditions when such links will be used for well-defined purposes. These linkages frequently involve tests that are intended to be equated and are therefore created to identical specifications; are highly similar in content emphasis, difficulty, and format; are equally reliable; and are expected to be administered under the same conditions. Equating different forms of college admissions tests by the College Board or the American College Testing Program are examples of this type of linkage. The committee is reviewing studies of this type of linkage because they may shed light on our charge, but we underscore that our conclusions do not apply to the linking of different forms of the same test.

There are other situations in which it is fairly routine for two tests to be linked and the results of the linkage to be used for well-defined purposes. For example, when a new test is introduced into a product line, a test publisher will establish links between the new product and the old one so that results obtained from the two tests can be compared. For example, CTB/McGraw Hill has linked the California Test of Basic Skills with its newer Terra Nova test; Harcourt Brace Educational Measurement has linked the Stanford Achievement Test 8 with the Stanford Achievement Test 9; Riverside Publishing has linked the Iowa Tests of Basic Skills M with earlier versions of the test. Sometimes the test specifications may have changed in response to shifts in educational emphases, and the old and new versions will not be as similar as two different versions of a test made to the same specifications; however, old and new versions can generally be successfully calibrated and put on the same scale.

Another routine use of linking occurs when states or schools change from one testing program to

another. In these cases it is not uncommon for a test publisher to conduct a study to link the two testing programs, even when the instruments were created by different publishers (Wendy Yen, personal communication). For example, when Virginia changed from the Iowa Tests of Basic Skills to the Stanford Achievement Test 9, the publisher of the Stanford 9, Harcourt Brace Educational Measurement, conducted a linkage study for Virginia that allowed trend lines for school and state data to be maintained. Such calibrations are not as robust as links of equivalent forms, but they suffice for comparing aggregate data.

The committee is also reviewing studies that describe state efforts to link their assessment results to NAEP, to estimate how schools or districts (but not individuals) might have performed if their students had participated in NAEP. There are also studies that compare trends on NAEP with trends on state tests, in order to evaluate states' progress against a national benchmark (Williams et al., 1995; Ercikan, 1997). Most recently, the National Center for Education Statistics of the Department of Education completed a study designed to link 4th and 8th grade mathematics and science results on NAEP and TIMSS; their aim was to estimate how groups of students who participated in the 1996 NAEP would have performed on the 1995 TIMSS (U.S. Department of Education, 1998). The committee is reviewing these studies (and others) in an effort to be comprehensive; we are aware, however, of fundamental differences between links across aggregates (states, districts) and links involving scores of individual students. For example, in linking to produce aggregate summary statistics for school districts or schools it is reasonable to incorporate important demographic information about the test takers into the linking function; such information would not be appropriate when reporting linked scores for individuals.

## DISTINCT CHARACTER OF NAEP

NAEP is a periodically administered, federally sponsored survey of a nationally representative sample of American students that assesses student achievement in key subjects. It combines the data from all test takers and uses the resulting aggregate information to monitor and report on the academic performance of U.S. children as a group, as well as by specific subgroups of the student population. NAEP was not designed to provide achievement information about individual students. Rather, NAEP reports the aggregate, or collective, performance of students in two ways—scale scores and achievement levels: the scale score results provide information about the distribution of student achievement for groups and subgroups in terms of a continuous scale; achievement levels are used to categorize student achievement as basic, proficient, or advanced (U.S. Department of Education, 1997).

NAEP makes use of a technique called matrix sampling, which enables it to achieve two goals. First, students are asked to answer a relatively small number of test questions so that the testing task given to students takes a relatively short time. Second, by asking different sets of questions of different students, the assessments cover a much larger array of questions than those given to any one student. NAEP's statistical design makes it possible to estimate the distribution of student scores by pooling data across subjects (Mislevy et al., 1992; Beaton and Gonzalez, 1995). The price paid for this flexibility is the inability of these assessments to collect enough data from any single student to allow valid individual scores to be reported.

NAEP's distinctive characteristics present special challenges of content comparability with other tests (e.g., Kenney and Silver, 1997). First, NAEP content is determined through a rigorous and lengthy consensus process that culminates in "frameworks" deemed relevant to NAEP's principal goal of monitoring aggregate student performance for the nation as a whole. NAEP content is not supposed to reflect particular state or local curricular goals, but rather a broad national consensus on what is or

should be taught; by design, its content is different from that of many state assessments (Campbell et al., 1994).

Second, NAEP's structure is unique: each student in the NAEP national sample takes only one booklet that contains a few short blocks of NAEP items in a single subject area (generally, three 15-minute or two 25-minute blocks), and no student's test booklet is fully representative of the entire NAEP assessment in that subject area. The scores for the blocks a student takes are used to predict his or her performance on the entire assessment. Thus, the portion of NAEP any one student takes is unlikely to be comparable in content to the full knowledge domain covered by an individual test taker in a state or commercial test (see, e.g., U.S. Department of Education, 1997; National Research Council, 1996; Beaton and Gonzalez, 1995; U.S. Congress, 1992). These issues greatly increase the difficulty of establishing valid and reliable links between commercial or state tests and NAEP.

# Findings

## COMPARABILITY: CONTENT, FORMAT, AND RELATED FEATURES

The content of a test is shaped by the kinds of knowledge and skills addressed in its questions ("items"). The committee's review indicates that content is not generally comparable among various state assessments and commercial tests, even when they are testing the same subjects. Middle-school mathematics, for instance, covers several subject areas of knowledge, such as arithmetic, algebra, and geometry: the content of one state's 8th grade mathematics test might focus largely on multiplication, division, and other number operations skills, while another test may stress pattern recognition and other pre-algebra skills (Bond and Jaeger, 1993). In reading, one 4th grade test may emphasize vocabulary and basic comprehension, while another may give greater weight to critical evaluation of an author's themes (Afflerbach et al., 1995).

A related content issue pertains to the skills and cognitive processes required to answer items. Off-the-shelf commercial tests and tests that are custom developed for states are increasingly constructed as mixed-model assessments that contain different types of items, including multiple-choice items and various kinds of open-ended questions for which students construct their own responses by filling in a blank, solving a problem, writing a short answer, writing a longer response, or completing a graph or diagram (see, e.g., Shavelson, Baxter, and Pine, 1992); Colorado, Connecticut, North Carolina, and Maryland are examples of states with mixed-model assessments. Some item types are very useful for testing student recall of factual material (a claim often made for certain types of multiple choice items); other item types are better suited to eliciting direct evidence of how well a student can solve problems.

The effect of format differences on linkages can be substantial. For example, the 1991 NAEP trial state assessment in mathematics contained both multiple choice and short-answer formats. Linn, Shepard, and Hartka (1992) found that when the two formats were scored separately, there was enough difference between the scores to change the rank order of the states in the mathematics assessment. For items with constructed responses (that is, not multiple choice), variations in scoring may also influence the validity of linkages because different scoring guides may credit different aspects of performance,

**BOX 3. Format Differences in Maryland**

The following description from Yen (1996:20) exemplifies the wide differences between two assessments of the same content area, administered to the same students in Maryland. They are the Maryland School Performance Assessment Program (MSPAP), a performance assessment used in high-stake school evaluations, and the Comprehensive Test of Basic Skills, Fourth Edition, or CTBS/4, published by CTB/McMillan McGraw-Hill in 1989.

> MSPAP and CTBS/4 differ in many ways. MSPAP is entirely performance based, and each student is given a limited number of reading selections or scenarios that require in-depth constructed and extended responses. In contrast, CTBS/4 samples a broader range of traditional objectives with a selected-response format and is a more indirect measure of student classroom performance. MSPAP is intended to "guide and goad" classroom instruction, while CTBS/4 is not intended as a model of instruction. MSPAP, which is targeted at raising student performance, contains many challenging items; CTBS/4 contains items that measure the full range of student performance. Each year three new forms of MSPAP are administered, with random assignment of forms to students; the same form of CTBS/4 is administered to all students every year. MSPAP results are used as part of a high-stakes program of evaluating schools; CTBS/4 results are part of the public reporting of school performance but are not included in the Maryland School Performance Index, which is used in school evaluations. Schools make individual decisions in terms of striking a balance between focusing on the material assessed with MSPAP and that assessed with CTBS/4.

even when the items appear similar (Linn, 1993). Issues such as how the scorers are trained and which scoring guidelines they use can affect the objectivity and consistency of scoring (Frederiksen and Collins, 1989). Some states, including Vermont and New Mexico, are trying out new assessment formats, such as systematically evaluating collections ("portfolios") of a student's work, that raise even more complex issues about comparability and scoring (Valencia and Au, 1997; Webb, 1995); see Box 3 for a discussion of format issues.

In short, content, format, and related issues are vitally important in linking, and existing commercially developed achievement tests and state assessments differ substantially among themselves and NAEP on these dimensions. The committee finds that the lack of strong comparability in these areas prevents the development of reliable and valid linkages. In addition, the committee finds that, in the cases that are germane to our concerns here, statistical linkages between tests with substantial differences in content and degrees of difficulty will not be accurate in the sense that they will not be consistent across subpopulations. This lack of consistency, a problem to which we return below, is directly due to the differences in content and test difficulty.

## DIVERSITY AND MULTIPLICITY OF TESTING PROGRAMS

Educational testing in the United States is diverse, reflecting the nation's history of state and local control over education policy. The number and variety of existing state and commercial tests pose formidable barriers to developing a single linking scale. State and commercial tests vary not only in content and format, but also in their target ages or grades, sampling techniques, policies for testing students with disabilities or with limited English proficiency, alignment with state and local curricula, score reporting procedures, and other factors (Bond et al., 1997).

Although commercially developed subject-matter achievement tests, especially the most widely used tests in U.S. schools, appear on the surface to be more similar than many existing state assessments, they, too, have significant differences that reflect the publishers' efforts to capture specialized

**BOX 4. California Comparability**

In 1996 California lawmakers determined that they wanted achievement information for monitoring school effectiveness, but, in the interest of respecting local control of educational issues, they did not want to mandate that all school districts use the same test. They therefore passed a bill encouraging school districts to choose achievement tests from a reviewed and approved list and then mandated the California Department of Education to develop a comparability scale that would allow lawmakers to accurately compare results from different assessments (see, Haertel, 1996; Wilson, 1996; Yen, 1996). Two different methodologies were explored in some depth. The first proposal suggested the development of a short list of acceptable commercial tests, any of which could be selected and administered by a local school district. These few tests would be linked in a manner similar to the Anchor Test Study (Loret et al., 1972). The second proposal was to develop a core reference test that comprehensively reflected California curriculum and to use that reference test as an anchor to which all other tests could be linked. In the end, the project was deemed too complex because more than 40 tests were submitted for comparison, and it was determined that it was too difficult to develop satisfactory links that would be stable over time. California decided to scrap the linkage proposal and selected one test to be administered in all of the state's schools.

markets and meet state and local demands for tests with particular features (Yen, 1998). The substantial variation that exists among commercially produced tests challenges the notion of selecting tests "off-the-shelf" and linking them: Box 4 illustrates an example of a recent attempt to link existing off-the-shelf tests and the difficulties that were encountered.

The complexity of linking even a small subset of existing tests could quickly render the task infeasible. For example, if the goal is to link just 15 different state assessments, it would be necessary to construct comparisons of more than 100 potential pairs of tests; each pair would require data collection, statistical analyses, and empirical validation (see also Los Angeles County Office of Education, 1997). An additional complicating factor could be the changing relationships between some of the tests**.** Frequent changes could necessitate continual updates to the development and validation of the equivalency scale (Loret et al., 1972; Linn, 1975; Wilson, 1996).

One might argue that pairwise comparisons are not necessary if all tests can be linked to a common scale, such as NAEP. Linking to NAEP simplifies the task in one respect by reducing the number of linkages that would have to be constructed. However, the design and purpose of NAEP complicates the task of linkage in another respect, and casts doubt on the validity of inferences that could be drawn from the link (see, e.g., McLaughlin, 1998). This is true because NAEP, by intent, does not produce scores for individuals and because individual students complete different parts of an entire NAEP assessment.

## STABILITY OF RESULTS

The testing landscape in the United States is not only diverse, but it is dynamic: states and districts have moved rapidly, especially during the last 10 years, to adopt new educational goals, new models of testing and assessment, and new strategies for aligning tests and assessments to state content standards (National Research Council, 1997).

Moreover, although there is some similarity and stability among the largest commercial testing programs, states that use commercial programs use them in very different ways. Many states have changed the design of their statewide assessments several times in the last decade and are continuing to do so. For example, some states are developing hybrids of commercial and state-developed tests or

customizing available off-the-shelf tests. Other states do not use commercial tests as part of their statewide assessment system (Roeber, Bond, and Braskamp, 1997). The diversity of the testing programs currently in the nation's schools is depicted in Table 1. (The committee realizes that information such as that tabulated here changes frequently, and may be summarized differently in different surveys. However, the main point is that the states' testing programs are extremely diverse in content, difficulty, and format (Jaeger, 1996).)

This continual change in educational goals and in the content of tests and assessments, which many people believe reflects a healthy dynamism in American education, makes linkage a moving target. Prior research has consistently shown that even if linkages between tests can be made at one time, they are difficult to maintain (Linn and Kiplinger, 1995). For example, suppose a link could be generated between a test in state A and another test in state B. Conducting the necessary analyses to establish the link takes time. It is quite possible that once the linkage methods are ready to be applied, one or both states will have changed their test format, content, or target group.

While NAEP does not change as frequently or as dramatically as state and commercial assessments, it, too, is not static. The content and nature of the NAEP instruments evolve gradually to reflect changing educational and assessment practices (National Research Council, 1996). These modifications in NAEP make it complicated to maintain stable linkages with state and commercial assessments, which are themselves evolving, and would minimize the validity of inferences from the linkages.

## TEST USES AND EFFECTS ON TEACHER AND STUDENT BEHAVIOR

Many states use assessments for multiple purposes related to educational improvement, such as program evaluation, curriculum planning, school performance reporting, and student diagnosis (U.S. Congress, 1992). More and more states are using (or are contemplating using) their assessment programs to make "high-stakes" decisions about people and programs, such as promoting students to the next grade, determining whether students will graduate from high school, grouping students for instructional purposes, making decisions about teacher tenure or bonuses, allocating resources to schools, or imposing sanctions on schools and districts (see, e.g., McLaughlin et al., 1995; McDonnell, 1997). Table 2 shows many of the varied uses of tests in our nation's schools today. (A companion report on appropriate test use will be issued by the National Research Council's Committee on the Fair and Appropriate Use of Educational Tests later this year.)

An important factor in testing goes under the heading of "stakes." When students are tested, various parties can have different concerns with, or stakes, in the outcomes. For example, a national survey of achievement, like NAEP, is a very low-stakes test for many of the parties concerned—the test takers, their parents, their teachers, and the district administrators. If NAEP is high-stakes for anyone, it is for policy makers who want to use NAEP data to assess the effectiveness of various educational reforms as they vary across the states or regions of the country. In contrast, tests that are used for high-school graduation or college admission are high-stakes for the students taking them and for their parents. Tests that are high stakes for the test takers affect their motivation to do their best. Tests that are high stakes for district administrators can result in various kinds of efforts to assist students in performing better than they would had the same test been of low stakes to those administrators.
These examples do not exhaust the possibilities of the effect of "stakes" on test results. When tests carrying different stakes for different parties are linked, one expects different linking functions to result than would be found if the stakes were similar.

Although forms of test-based educational accountability vary across states and districts, changes in how tests are used inevitably lead to changes in how teachers and students react to them (Koretz,

TABLE 1 State Testing: A Snapshot of Diversity

| State | Use of Commercial Tests | Use of Other Assessments |
|---|---|---|
| Alabama | Stanford Achievement Test 9, Otis Lenin School Ability Test | Alabama Kindergarten Assessment, Alabama Direct Assessment of Writing, Differential Aptitude Test, Basic Competency Test, Career Interest Inventory, End-of-Course Algebra and Geometry Test, Alabama High School Basic Skills Exit Exam |
| Alaska | California Achievement Test 5 | |
| Arizona | Stanford Achievement Test 9 | |
| Arkansas | Stanford Achievement Test 9 | High School Proficiency Test |
| California | Stanford Achievement Test 9 | Golden State Examinations |
| Colorado | Custom developed | CTB item banks, NAEP items, and state items |
| Connecticut | Custom developed | Connecticut Mastery Test, Connecticut Academic Performance Test |
| Delaware | Custom developed | State-developed writing assessment |
| Florida | Custom developed | High School Competency Test, Florida Writing Assessment Program |
| Georgia | Iowa Test of Basic Skills, Test of Achievement Proficiency | Curriculum-based Assessments, Georgia High School Graduation Tests, Georgia Kindergarten Assessment Program, Writing Assessment |
| Hawaii | Stanford Achievement Test 8 | Hawaii State Test of Essential Competencies, Credit by Examination |
| Idaho | Iowa Test of Basic Skills Form K, Test of Achievement Proficiency | Direct Writing Assessment, Direct Mathematics Assessment |
| Illinois | Custom developed | Illinois Goals Assessment Program |
| Indiana | Custom developed | Indiana StatewideTesting for Educational Progress Plus |
| Iowa | No mandated statewide testing program, approximately 99 percent of all districts participate in the Iowa Test of Basic Skills on a voluntary basis | |
| Kansas | Custom developed | Kansas Assessment Program (Kansas University Center for Educational Testing and Evaluation) |
| Kentucky | Custom developed | Kentucky Instructional Results Information System |
| Louisana | California Achievement Test 5 | Louisiana Educational Assessment Program |
| Maine | Custom developed | Maine Educational Assessment (Advanced Systems in Measurement Inc.) |
| Maryland | Custom developed, Comprehensive Test of Basic Skills 5 | Maryland Student Performance Assessment Program, Maryland Functional Tests, Maryland Writing Test |
| Massachusetts | Iowa Test of Basic Skills, Iowa Test of Educational Development | |
| Michigan | Custom developed | Michigan Educational Assessment Program: Criterion-referenced tests of 4th-, 7th-, and 11th-grade students in mathematics and reading and 5th-, 8th-, and 11th-grade students in science and writing; Michigan High School Proficiency Test |
| Minnesota | Custom developed | 96-97 students took minimum competency literacy tests in reading and mathematics |
| Mississippi | Iowa Test of Basic Skills, Test of Achievement Proficiency | Functional Literacy Examination, Subject Area Testing Program |

*continued*

TABLE 1 (Continued)

| State | Use of Commercial Tests | Use of Other Assessments |
|---|---|---|
| Missouri | Custom developed, TerraNova | Missouri Mastery and Achievement Test |
| Montana | Stanford Achievement Test, Iowa Test of Basic Skills, Comprehensive Test of Basic Skills | |
| Nebraska | No statewide assessment program in 96-97 | |
| Nevada | TerraNova | Grade 8 Writing Proficiency Exam, Grade 11 Proficiency Exam |
| New Hampshire | Custom developed | New Hampshire Education Improvement and Assessment Program (Advanced Systems in Measurement and Evaluation, Inc.) |
| New Jersey | Custom developed | Grade 11 High School Proficiency Test, Grade 8 Early Warning Test |
| New Mexico | Iowa Test of Basic Skills, Form K | New Mexico High School Competency Exam, Portfolio Writing Assessment, Reading Assessment for Grades 1 and 2 |
| New York | Custom developed | Occupational Education Proficiency Examinations, Preliminary Competency Tests, Program Evaluation Tests, Pupil Evaluation Program Tests, Regents Competency Tests, Regents Examination Program, Second Language Proficiency Examinations |
| North Carolina | Iowa Test of Basic Skills | North Carolina End of Grade |
| North Dakota | Comprehensive Test of Basic Skills/4, TCS | |
| Ohio | Custom developed | Fourth-, Sixth-, Ninth-, and Twelfth-Grade Proficiency Tests |
| Oklahoma | Iowa Test of Basic Skills | Oklahoma Core Curriculum Tests |
| Oregon | Custom developed | Reading, Writing, and Mathematics Assessment |
| Pennsylvania | Custom developed | Writing, Reading, and Mathematics Assessment |
| Rhode Island | Metropolitan Achievement Test 7, Custom developed | Health Performance Assessment, Mathematics Performance Assessment, Writing Performance Assessment |
| South Carolina | Metropolitan Achievement Test 7, Custom developed | Basic Skills Assessment Program |
| South Dakota | Stanford Achievement Test 9, Metropolitan Achievement Test 7 | |
| Tennesee | Custom developed | Tennessee Comprehensive Assessment Program (TCAP) Achievement Test Grades 2-8, TCAP Competency Graduation Test, TCAP Writing Assessment Grades 4, 8, and 11 |
| Texas | Custom developed | Texas Assessment of Academic Skills, Texas End-of-Course Test |
| Utah | Stanford Achievement Test 9, Custom developed | Core Curriculum Assessment Program |
| Vermont | Has a voluntary state assessment program | New Standards reference exams in math, Portfolio assessment in math and writing |
| Virginia | Customized off the shelf | Literacy Passport Test, Degrees of Reading Power, Standards of Learning Assessments, Virginia State Assessment Program |

TABLE 1 (Continued)

| State | Use of Commercial Tests | Use of Other Assessments |
|---|---|---|
| Washington | Comprehensive Test of Basic Skills 4, Curriculum Frameworks Assessment System | |
| West Virginia | Comprehensive Test of Basic Skills | Writing Assessment, Metropolitan Readiness Test |
| Wisconsin | TerraNova, Custom developed | Knowledge and Concepts Tests, Wisconsin Reading Comprehension Test at Grade 3 |
| Wyoming | State assessment program in vocational education only for students grades 9-12 | |

NOTES: Custom developed assessments result from a joint venture between a state and a commercial test publisher to design a test to the state's specification, perhaps to more closely match the state's curriculum than an off-the-shelf test does. Customized off-the-shelf assessments result from modifications to a commerical test publisher's existing product.

SOURCE: Data from 1997 Council of Chief State School Officers Fall State Student Assessment Program Survey

1998). Indeed, one of the underlying rationales for test-based accountability is to spur changes in teaching and learning. These uses are hotly debated and beyond the scope of this report (Jones, 1997). For our purposes, it is sufficient to note that the difficulty of maintaining linkages between tests is exacerbated when test results have significant consequences for individuals or schools.

In these situations, teachers may change what and how they teach to help students respond to the content and problems on the test (Shepard and Dougherty, 1991), schools and districts may align curriculum more closely with test content, and test takers may have stronger motivation to do well (e.g., Koretz et al., 1991). Performance gains on tests used for accountability (high-stakes tests) will often not be reflected in scores on tests used for monitoring or other non-accountability (low-stakes) purposes. The resulting differences in student performance could alter the relationship between linked tests (Shepard et al., 1996; Yen, 1996). Hence, any valid linkages created initially would have to be reestablished regularly, which would raise important questions about any hoped-for cost-effective advantages of linkage.

The effects of test use on student and teacher behavior pose a special problem for linkage with NAEP. To protect its historical purpose as a monitor of educational progress, NAEP was designed expressly with safeguards to prevent it from becoming a high-stakes test. As a result, the motivation level of students who participate in NAEP may be low (O'Neil et al., 1992; Kiplinger and Linn, 1996), and they may not always exhibit peak performance. Linkages between a low-stakes instrument like NAEP and high-stakes state or commercial tests are likely to be misleading because students are likely to put forth more effort for the latter kinds of tests than for the former.

## POPULATION OR SUBGROUP DIFFERENCES

When the function that links Test A with Test B differs for different groups, for example, boys and girls, it does not indicate that one group is "better" than the other. Rather, it means that a boy and a girl with the same score on Test A would be expected to have different scores on Test B, and that this

TABLE 2 Student Testing: Diversity of Purpose

| State | Decisions for Students | Decisions for Schools | Instructional Purposes |
|---|---|---|---|
| Alabama | High school graduation | School performance reporting | Student diagnosis or placement; improve instruction; program evaluation |
| Alaska | | School performance reporting | Improve instruction |
| Arizona | | School performance reporting | Student diagnosis or placement; improve instruction; program evaluation |
| Arkansas | | School performance reporting | Student diagnosis or placement; improve instruction; program evaluation |
| California | Student diagnosis or placement | | Student diagnosis or placement |
| Colorado[a] | | | |
| Connecticut | Student diagnosis or placement | Awards or recognition; school performance reporting | Student diagnosis or placement; improve instruction; program evaluation |
| Delaware | | | Student diagnosis or placement; improve instruction; program evaluation |
| Florida | High school graduation | | Improve instruction; program evaluation |
| Georgia | High school graduation | School performance reporting | Student diagnosis or placement; improve instruction; program evaluation |
| Hawaii | High school graduation | Awards or recognition; school performance reporting | Student diagnosis or placement; improve instruction; program evaluation |
| Idaho | | School performance reporting | Improve instruction |
| Iowa[a] | | | |
| Illinois | | Accreditation | |
| Indiana | | Awards or recognition; school performance reporting | Student diagnosis or placement; improve instruction; program evaluation |
| Kansas | | School performance; reporting; accreditation | Student diagnosis or placement; improve instruction; program evaluation |
| Kentucky | | Awards or recognition | Improve instruction; program evaluation |
| Louisiana | Student promotion; high school graduation | Awards or recognition; school performance reporting | Student diagnosis or placement; improve instruction; program evaluation |
| Maine | Student diagnosis or placement | | Improve instruction; program evaluation |

TABLE 2 (Continued)

| State | Decisions for Students | Decisions for Schools | Instructional Purposes |
|---|---|---|---|
| Maryland | High school graduation | School performance reporting; skills guarantee; accreditation | Student diagnosis or placement; improve instruction; program evaluation |
| Massachusetts | | School performance reporting | Improve instruction |
| Michigan | Student diagnosis or placement; endorsed diploma | Awards or recognition; school performance reporting; accreditation | Improve instruction; program evaluation |
| Minnesota[a] | | | |
| Mississippi | High school graduation | School performance reporting; skills guarantee; accreditation | Student diagnosis or placement; improve instruction; program evaluation |
| Missouri | | School performance reporting; accreditation | Improve instruction; program evaluation |
| Montana | | | Improve instruction; program evaluation |
| Nebraska[a] | | | |
| Nevada | High school graduation | School performance reporting; accreditation | Improve instruction; program evaluation |
| New Hampshire | | | Improve instruction; program evaluation |
| New Jersey | High school graduation | School performance reporting; accreditation | Student diagnosis or placement; improve instruction |
| New Mexico | High school graduation | School performance reporting; accreditation | Student diagnosis or placement; improve instruction; program evaluation |
| New York | Student diagnosis or placement; student promotion; honors diploma; endorsed diploma; high school graduation | School performance reporting | Improve instruction; program evaluation |
| North Carolina | Student diagnosis or placement; student Promotion; high school graduation | | Improve instruction; program evaluation |
| North Dakota | Student diagnosis or placement | | Student diagnosis or placement; improve instruction; program evaluation |
| Ohio | High school graduation | Awards or recognition; school performance reporting | Improve instruction; program evaluation |
| Oklahoma | | School performance reporting; accreditation | Student diagnosis or placement; improve instruction; program evaluation |

*continued*

TABLE 2 (Continued)

| State | Decisions for Students | Decisions for Schools | Instructional Purposes |
|---|---|---|---|
| Oregon | | School performance reporting | Improve instruction; program evaluation |
| Pennsylvania | | School performance reporting | Student diagnosis or placement; program evaluation |
| Rhode Island | | School performance reporting | Improve instruction; program evaluation |
| South Carolina | Student promotion; high school graduation | Awards or recognition; school performance reporting; skills guarantee | Student diagnosis or placement; improve instruction; program evaluation |
| South Dakota | | | Improve instruction; program evaluation |
| Tennessee | Endorsed diploma; high school graduation | | Student diagnosis or placement; improve instruction; program evaluation |
| Texas | Student diagnosis or placement; high school graduation | | Student diagnosis or placement; improve instruction; program evaluation |
| Utah | Student diagnosis or placement | School performance reporting | Student diagnosis or placement; improve instruction; program evaluation |
| Vermont | | School performance reporting | Student diagnosis or placement; improve instruction; program evaluation |
| Virginia | Student diagnosis or placement; student promotion; high school graduation | School performance reporting | Student diagnosis or placement; improve instruction; program evaluation |
| Washington | | School performance reporting | Student diagnosis or placement; improve instruction; program evaluation |
| West Virginia | | Skills guarantee; Accreditation | Improve instruction |
| Wisconsin | | School performance reporting | Program evaluation |
| Wyoming | | | Improve instruction; program evaluation |

[a]Colorado, Minnesota, and Nebraska did not administer any statewide assessments in 1995-96. Iowa does not administer a statewide assessment.

SOURCE: Data from 1996 Council of Chief State School Officers Fall State Student Assessment Program Survey.

effect is consistent for members of the two groups. Researchers generally suppose that group differences occur because of differing test content or format, different motivation levels, or differences in prior exposure to relevant learning opportunities. Perhaps the material in Test A is more familiar to one group than the other, while the material in Test B is equally familiar to both groups. Alternatively, one group might be motivated to perform well on one test, while both groups were equally motivated on Test B.

Simply put, it is often the case that the relative differences among the test performances of different groups of students will vary from test to test, depending on a host of factors that are subtle but important. For example, on mathematics tests boys may do better on word problems while girls may do better solving equations. When this is true, overall estimates of gender differences in 8th grade mathematics performance will depend on the relative emphasis a test gives to these two areas. Unless the two tests are very closely aligned in content, linking them might require separate formulas for boys and girls because a single linking formula would underestimate performance for one group and overestimate it for the other. Another example is that student achievement on two tests with differing emphases on algebra could vary widely across the states as a function of when and to what extent algebra is introduced into the middle school curriculum. As a result, students from different states who obtain the same score on one test (e.g., a commercial test) might have different estimated (linked) scores on a second test, such as NAEP (e.g., McLaughlin, 1998). These problems are attributable in part to the tests themselves, but linkage magnifies them and increases the risk of unfair inferences about individual achievement.

## REPORTING RESULTS IN TERMS OF NAEP ACHIEVEMENT LEVELS

Linking other tests to NAEP raises the possibility of reporting individual student scores on state and commercial tests in terms of the NAEP achievement levels. The committee explored this issue and finds that such links would raise new and significant methodological problems (see Wu, Royal, and McLaughlin, 1997).

To understand them, one must recognize that all test scores carry with them some amount of uncertainty or "noise," an issue usually treated in the testing literature under the heading "reliability" (see, e.g., Feldt and Brennan, 1989; American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1985). Because scores are less than 100 percent reliable, it is never possible to assign students to achievement levels with complete certainty (Johnson and Mazzeo, 1998). The two key issues are the likelihood that students will be misclassified and the degree of error in the classification. Clearly, the more reliable the test, the less ambiguity there will be in the assignment of students to categories of performance. Unfortunately, the empirical evidence that the committee has reviewed to date suggests that transforming performances on selected existing assessments to the NAEP achievement levels produces results with substantial practical ambiguity.

For example, consider a 4th grade student with a reasonably good score on a state or commercial reading test. Transforming this child's score into the NAEP achievement levels could easily produce the following type of report: "Sally scored [x] on the [State Reading Assessment]. Of 100 students with the same score, 10 are likely to be in the 'below basic' category; 60 are likely to be 'basic;' 28 are likely to be 'proficient;' and 2 are likely to be in the highest, or 'advanced,' category." Alternatively, the report could be issued in terms of Sally's probabilities of falling in the various categories (Johnson and Mazzeo, 1998). This ambiguity will be due to measurement error in the student's score on the state assessment; to measurement error in NAEP; to the less than perfect correlation between the state

assessment scores and NAEP scores; to potential differences in linking functions by different subgroups; and to other unidentified sources of measurement error.

The committee has not been able to conduct a thorough study of parental and public reaction to this kind of scenario, but we caution that one of the more important putative purposes of linkage—providing clear and relevant information about the performance of individual students—might be severely undermined by the need to report information which, in order to be faithful to the underlying statistics, must be ambiguous in its meaning.

Table 3 presents a summary of some major studies of linkage between different tests and assessments. Some of the studies describe methods for linking to NAEP and the implications for such a link.

TABLE 3 Abridged Summaries of Prior Linkage Research

| Study | Purpose | Methodology | Key Findings |
|---|---|---|---|
| The Anchor Test Study (Loret et al., 1972, 1973) | To develop an equivalency scale to compare reading test results for Title I program evaluation.<br>The study was sponsored by a $1,000,000 contract with the U.S. Office of Education. | Number of participants: 200,000 students for norming phase; 21 sample groups of approximately 5,000 students each for the equating phase.<br>Eight tests, representing almost 90% of reading tests being administered in the states at that time, were selected for the study.<br>Participants took two tests.<br>Created new national norms for one test and through equating, all eight tests.<br>Administered different combinations of standardized reading tests to different subjects taking into account the need to balance demographic factors and instructional differences. | Tests with similar content can be linked together with reasonable accuracy.<br>Relationships between tests were determined to be reasonably similar for male and female students but not for racial groups.<br>The equivalency scale was accurate for individuals but aggregated results, e.g., school or district, would have increased error stemming from combining results.<br>Every time a new test is introduced the procedure has to be replicated for that test.<br>The stability of the linkage has to be reestablished regularly because instruction on one test but not on others can invalidate the linkage. |
| Projecting to the NAEP Scale: Results from the North Carolina End-of-Grade Testing Program (Williams et al., 1995) | To link a comprehensive state achievement test to the NAEP scale for mathematics so that the more frequently administered state tests could be used for purposes of monitoring progress of North Carolina students with respect to national achievement standards. | A total of 2,824 students from 99 schools was tested using 78 items from a short form of the North Carolina End-of-Grade Test and two blocks of released 1992 NAEP items that were embedded in the test.<br>Test booklets were spiraled so that some students took NAEP items first, others took North Carolina End-of-Grade items first.<br>The final linkage to the NAEP scale used projection. Scores from the NAEP blocks were determined from student responses using NAEP parameters but not the conditioning analysis used by NAEP. Regular scores from the North Carolina test were used. | A satisfactory linking was obtained for statewide statistics as a whole that were accurate enough to predict NAEP means or quartile distributions with only modest error.<br>The linkages had to be adjusted separately for different ethnic groups, demonstrating that the linking was inappropriate for predicting individual scores from the North Carolina test to the NAEP scale.<br>The following were considered important factors in establishing a strong link: content on the North Carolina test was closely aligned with state curriculum and NAEP's was not; student performance was affected by the order of the items in their test booklets; motivation or fatigue affects performance for some students. |

*continued*

TABLE 3 (Continued)

| Study | Purpose | Methodology | Key Findings |
|---|---|---|---|
| Linking Statewide Tests to NAEP (Ercikan, 1997) | To examine the accuracy of linking statewide test results to NAEP by comparing the results of four states' assessment programs with the NAEP results for those states. | Compared each state's assessment data to its NAEP data using equipercentile comparisons of score distributions. Since none of the four states used exactly the same form of the California Achievement Test for their state testing program, state results had to be converted to a common scale. This scale was developed by the publisher of the California Achievement Test series. | The link from separate tests to NAEP varies from one state to the next. It was not possible to determine whether the state-to-state differences were due to the different test(s), the moderate content alignment, the motivation of the students, or the nature of the student population.<br>Linking state tests to NAEP (by matching distributions) is so imprecise that results should not be used for high-stakes purposes. |
| Toward World-Class Standards: A Research Study Linking International and National Assessments (Pashley and Phillips et al., 1993) | To pilot test a method for obtaining accurate links between the International Assessment of Educational Progress (IAEP) and NAEP so that other countries can be compared with the United States, both nationally and at the state level, in terms of NAEP performance standards. | A sample of 1,609 U.S. grade eight students were assessed with both IAEP and NAEP instruments in 1992 to establish a link between these assessments.<br>Based on test results from the sample testing, the relationships between IAEP and NAEP proficiency estimates were investigated.<br>Projection methodology was used to estimate the percentages of students from the 20 countries, assessed with the IAEP, who could perform at or above the three performance levels established for NAEP.<br>Various sources of statistical error were assessed. | The methods researchers use to establish links between tests (at least partially) determine how valid the link is for drawing particular inferences about performance.<br>Establishing this link required a linking sample of students who took both assessments.<br>While it is possible to establish an accurate statistical link between the IAEP and NAEP assessments, policy makers, among others, should proceed with caution when interpreting results from such a link.<br>The fact that the IAEP and NAEP were fairly similar in construction and scoring made the linking easier.<br>The effects of unexplored sources of non-statistical error, such as motivation levels, had on the results was not determined. |

TABLE 3 (Continued)

| Study | Purpose | Methodology | Key Findings |
|---|---|---|---|
| Comparing the NAEP Trial State Assessment (TSA) Results with the IAEP International Results (Beaton and Gonzalez, 1993) | To determine how American students compare to foreign students in mathematics, and how well foreign students meet the NAGB mathematics standards. | At that time data were not available for examinees that took both assessments, therefore they relied on a simple distribution-matching procedure.<br>Rescaled scores to produce a common mean and standard deviation on the two tests.<br>Translated IAEP scores into NAEP scores by aligning the means and standard deviations for the two tests.<br>Transformed the IAEP scores for students in the IAEP samples in each participating country into equivalent NAEP scores. | Moderation procedures are sensitive to age/grade differences.<br>The IAEP and NAEP have many similarities but are not identical and differ in some significant ways.<br>Results of the linking were different for countries with high average IAEP scores.<br>Different methods of linking the IAEP and NAEP can produce different results, and further study is necessary to determine which method is best. |
| Linking to a Large-Scale Assessment: An Empirical Evaluation (Bloxom et al., 1995) | To compare the mathematics achievement of new military recruits with the general U.S. student population, using a link between the Armed Services Vocational Aptitude Battery (ASVAB) and NAEP. The emphasis of the study was to provide and illustrate an approach for empirically evaluating the statistical accuracy of such a linkage, | A sample of 8,239 applicants for military service was administered an operational ASVAB and an NAEP survey in 1992. These applicants were told that there were no stakes attached to the NAEP survey.<br>ASVAB scores were projected on the NAEP scale in mathematics to allow for comparison between the achievement of military applicants with the general U.S. population of 12th grade students.<br>Statistical checks were made by constructing the link separately for low-scoring candidates and for high-scoring candidates. | Statistically, an accurate distribution of recruit achievement can be found by projecting onto the NAEP scale.<br>Factors related to motivation may have underestimated the assessment-based proficiency distribution of recruits in this study, meaning that in spite of the statistical precision of the linkage, the resulting estimates may not be valid for practical purposes. |

*continued*

TABLE 3 (Continued)

| Study | Purpose | Methodology | Key Findings |
|---|---|---|---|
| The Potential of Criterion-Referenced Tests with Projected Norms (Behuniak and Tucker, 1992) | To determine if norm-referenced scores could be provided, for the purpose of Chapter 1 program evaluation, by linking the Connecticut Mastery Test, a criterion-referenced test closely aligned with state curriculum, and a national "off-the-shelf" norm-referenced achievement test. The purpose of the linking was to meet Federal guidelines for Chapter 1 reporting without requiring students to take two tests. | Compared two tests, the Metropolitan Achievement Test 6 (MAT 6) and the Stanford Achievement Test 7 (SAT 7) to determine which was more closely aligned with state content standards. Selected the MAT 6 for the study.<br>For a relevant population, calibrated the items from the two instruments in a given subject as a single IRT calibration then used the results to calibrate the tests.<br>Linked results using equipercentile equating.<br>Examined changes over two years to check the stability of the link. | There were enough content differences between the two norm-referenced tests and the Connecticut Mastery Test to decide that one test would make a better, if not perfect, candidate for linking to the state test than the other.<br>It was possible to develop a link between the MAT 6 and the Connecticut Mastery Test that accurately predicted Normal Curve Equivalent scores for the MAT 6 from the CMT but no good validity checks were used.<br>The linking function changed somewhat over time and the authors believed that this divergence would continue because teachers were gearing instruction to state standards which were more closely aligned with the Connecticut test than the Metropolitan. Thus, the linking would have to be reestablished regularly to remain valid for the purposes that it was intended to serve. |

TABLE 3 (Continued)

| Study | Purpose | Methodology | Key Findings |
|---|---|---|---|
| Linking Statewide Tests to the National Assessment of Educational Progress: Stability of Results (Linn and Kiplinger, 1995) | To investigate the adequacy of linking statewide standardized test results to the National Assessment of Educational Progress (NAEP) to allow for accurate comparisons between state academic performance and the national performance levels measured by NAEP. | Obtained two years (1990 and 1992) results from four states' testing programs and corresponding results from the NAEP-TSA for the same two years. (Standardized tests used in the four states were different.)<br>Used equipercentile equating procedures to compare data from state tests and NAEP.<br>The standardized test results were converted to the NAEP scale using the 1990 data and resulting conversion tables were then applied to the 1992 data.<br>Examined content match between standardized tests and NAEP and re-analyzed data using subsections of the standardized tests and NAEP. | The link could estimate average state performance on NAEP, but was not accurate for scores at the top or bottom of the scale.<br>The equating function diverged for males and females, meaning that NAEP scores for a state would have been over predicted if the equating function for males was used rather than the equating function for females.<br>Linking standardized tests to NAEP using equipercentile equating procedures is not sufficiently trustworthy to use for other than rough approximations.<br>Designing tests in accordance with a common framework might make linking more feasible. |

*continued*

TABLE 3 (Continued)

| Study | Purpose | Methodology | Key Findings |
|---|---|---|---|
| Using Performance Standards to Link Statewide Achievement Results to NAEP (Waltman, 1997) | To investigate how the comparability of performance standards obtained by using both statistical and social moderation to link NAEP standards to the ITBS. | Compared 1992 NAEP-TSA with ITBS for Iowa 4th grade public school students.<br>Used two different types of linking for separate facets of the study.<br>A socially moderated linkage was obtained by setting standards independently on the ITBS using the same achievement-level descriptions used to set the NAEP achievement levels.<br>An equipercentile procedure was used to establish a statistically moderated link. | For students who took both assessments, the corresponding achievement regions on the NAEP and ITBS scales produced low to moderate percents of agreement in student classification. Agreement was particularly low for students at the advanced level, two-thirds or more were classified differently.<br>Cut-scores on the ITBS scale, established by moderation, were lower than those used by NAEP, resulting in more students being classified as basic, proficient, or advanced on the ITBS than estimated by NAEP, possibly due to content and skills-standards mismatch between the ITBS and NAEP.<br>The equipercentile linkage was reasonably invariant across types of communities, in terms of percentages of students classified at each level.<br>Regardless of the method used to establish the ITBS cut-scores or the criteria used to classify students, the inconsistency of student-level match limits even many inferences about group performances. |

TABLE 3 (Continued)

| Study | Purpose | Methodology | Key Findings |
|---|---|---|---|
| Study of the Linkages of 1996 NAEP and State Mathematics Assessments in Four States (McLaughlin, 1998) | To address the need for clear, rigorous standards for linkage; to provide the foundation for developing practical guidelines for states to use in linking state assessments to NAEP; and to demonstrate that it is important for educational policy makers to be aware that linkages that support one use may not be valid for another. | A sample of four states that had participated in the 1996 State NAEP mathematics assessment and whose state assessment mathematics tests could potentially be linked to NAEP at the individual student level participated in this study. Participating states used different assessments in their state testing programs.<br>There were eight linkage samples, ranging in size from 1,852 to 2,444 students.<br>Study matched students who participated in the NAEP assessment in their states with their scores on the state assessment instrument using projection with multilevel regression. | Links were not sufficiently accurate to permit reporting individual student proficiency on NAEP based on the state assessment score.<br>Links differed noticeably by minority status and school district, in all four states. Students with the same state assessment score would be projected to have different standings on the NAEP proficiency scale, depending on their minority status and school district. |
| The Maryland School Performance Program: Performance Assessment with Psychometric Quality Suitable for High Stakes Usage (Yen and Ferrarra, 1997) | To compare the Maryland State Performance Assessment (MSPAP) with the California Test of Basic Skills (CTBS) in order to establish the validity of the state test in reference to national norms. | Compared results from a group of 5$^{th}$ grade students who took both the MSPAP and the CTBS—correlations were obtained.<br>The intent was to establish the validity of the MSPAP so a link was not obtained. | Intercorrelations of the two tests indicated that the two measures were assessing somewhat different aspects of achievement. |

*continued*

TABLE 3 (Continued)

| Study | Purpose | Methodology | Key Findings |
|---|---|---|---|
| A TIMSS-NAEP Link (Johnson, 1998) | To provide useful information about the performance of states relative to other countries. The study broadly compares state eighth-grade mathematics and science performance for each of 44 states and jurisdictions participating in the NAEP with the 41 nations who participated in TIMSS. | The study provides predicted TIMSS results for 44 states and jurisdictions, based on their actual NAEP results. A statistically moderated link was used to establish the link between NAEP and TIMSS based on applying formal linear equating procedures. The link was established using reported results from the 1995 administration of TIMSS in the U.S. and the 1996 NAEP and matching characteristics of the score distributions for the two assessments. Validated the linking functions using data provided by states that participated in both state-level NAEP and state-level TIMSS but were not included in the development of the original linking function. | Although all of the findings have not yet been released, apparently some links were satisfactory and others were not. |

# Conclusions

Currently administered state and commercial achievement tests and NAEP vary significantly in terms of their content emphasis, types and difficulty of test questions, and the thought processes they require of students. In addition, these tests vary substantially in how and when they are administered, whether all students respond to the same sets of questions, how closely the tests are related to what is taught in school, how they are scored, and how the scores are reported and used (Roeber et al., 1997).

Therefore, the committee concludes that:

**Comparing the full array of currently administered commercial and state achievement tests to one another, through the development of a single equivalency or linking scale, is not feasible.**

**Reporting individual student scores from the full array of state and commercial achievement tests on the NAEP scale, and transforming individual scores on these various tests and assessments into the NAEP achievement levels, is not feasible.**

Although the committee concludes that it is not feasible to link the full array of existing tests to each other or to NAEP, it is exploring issues involved in developing linkages between specified subsets of these tests. Among the questions to be considered in our final report are whether criteria might be developed to help evaluate the quality of proposed linkages between various tests, what research would be required to develop such criteria, and what would be the longer term policy implications of selecting some of the many tests used by states and localities for linkage to a common scale or NAEP or both.

# References


Achieve, Inc.
1998 About Achieve. Available electronically at http://www.achieve.org, [March 1].

Afflerbach, P.A.
1995 Content validation of the 1994 NAEP in reading: Classifying items according to the reading framework. In *Assessment in Transition: 1994 Trial State Assessment Report on Reading: Background Studies*, R. Linn, R. Glaser, and G. Bohrnstedt, eds. Stanford, CA: The National Academy of Education.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education
1985 *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association, American Psychological Association, and National Council on Measurement in Education.

Beaton, A.E., and E.J. Gonzalez
1993 Comparing the NAEP trial assessment results with the IAEP international results. In *The Trial State Assessment: Prospects and Realities: Background Studies 1993*. Washington, DC: National Academy of Education.
1995 *The NAEP Primer*. Center for the Study of Testing, Evaluation, and Educational Policy, Chestnut Hill, MA: Boston College.

Behuniak, Peter, and Charlene Tucker
1992 The potential of criterion-referenced tests with projected norms. *Applied Measurement in Education* 5(4):337-353.

Bloxom, Bruce, Peter J. Pashley, W. Alan Nicewander, and Duanli Yan
1995 Linking to a large-scale assessment: An empirical evaluation. *Journal of Educational and Behavioral Statistics* 20 (Spring):1-26.

Bond, Linda A.
1995 *Norm-Referenced Testing and Criterion-Referenced Testing: The Difference in Purpose, Content, and Interpretation of Results*. Oak Brook, IL: North Central Regional Educational Laboratory.

Bond, Linda, Edward Roeber, and David Braskamp
1997 *Trends in State Student Assessment Programs: Fall 1996*. Washington, DC: Council of Chief State School Officers.

Bond, Lloyd, and Richard M. Jaeger
1993 *Judged Congruence Between Various State Assessment Tests in Mathematics and the 1990 National Assessment of Educational Progress Item Pool for Grade-8 Mathematics*. Center for Educational Research and Evaluation. Greensboro NC: University of North Carolina.

Campbell, J.R., P.L. Donahue, C.M. Reese, and G.W. Phillips
1994 *NAEP 1994 Reading Report Card for the Nation and the States*. Office of Educational Research and Improvement, National Center for Educational Statistics. Washington, DC: U.S. Department of Education.

Ercikan, Kadriye
1997 Linking statewide tests to the NAEP: Accuracy of combining test results across states. *Applied Measurement in Education* 10(2):145-159.

Feldt, L.S., and R.L. Brennan
1989 Reliability. In *Educational Measurement,* Third Edition, R.L. Linn, ed. New York: MacMillian Publishing Company.

Frederiksen, J.R., and A. Collins
1989 A systems approach to educational testing. *Educational Researcher* 18(9):27-32.

Haertel, E.H.
1996 Test linking and comparability. In *Proceedings from the California Comparability Symposium, September 1997*. Los Angeles, CA: County Office of Education.

Holland, Paul W., and D.B. Rubin, eds.
1982 *Test Equating*. New York: Academic Press.

Jaeger, Richard M.
1996 Content Congruence as a Factor in the Linking of State Assessments to NAEP. Paper presented at the Council of Chief State School Officers Large-Scale Assessment Conference June 23-26, Phoenix, AZ. University of North Carolina, Greensboro.

Johnson, Eugene
1998 *Linking the National Assessment of Educational Progress and the Third International Mathematics and Science Study for Eighth Grade: A Research Report*. National Center for Education Statistics, Publication No. NCES 98-499. Washington, DC: U.S. Department of Education.

Johnson, E.G., and J. Mazzeo
1998 Linking the Voluntary National Test with NAEP and TIMSS. Paper presented at the NCME Invited Session on Measurement Issues on the Voluntary National Tests, San Diego, CA. Educational Testing Service, Princeton, NJ.

Jones, Lyle V.
1997 National Tests and Educational Reform: Are They Compatible? Policy Information Center, Educational Testing Service, Princeton, NJ.

Kenney, Patricia Ann, and Edward Silver
1997 Content Analysis Project — State and NAEP Mathematics Assessment. Proposal Summary. Learning Research and Development Center, University of Pittsburgh.

Kiplinger, V.L., and R.L. Linn
1996 Raising the stakes of test administration: The impact on student performance on the National Assessment of Educational Progress. *Educational Assessment* 3(2):111-133.

Kolen, Michael J., and Robert L. Brennan
1995 *Test Equating: Methods and Practices*. New York: Springer-Verlag.

Koretz, Daniel
1998 Evidence Pertaining to the Validity of Score Gains on the Kentucky Instructional Results Information System (KIRIS). Paper presented at the symposium: Establishing Meaning: Validity Evidence for the Kentucky Instructional Results Information System (KIRIS), at the annual meeting of the National Council on Measurement in Education, April 16, San Diego, CA. RAND, Washington, DC.

Koretz, Daniel, R.L. Linn, S. Dunbar, and L. Shepard
1991 The Effects of High-Stakes Testing on Achievement: Preliminary Findings About Generalization Across Tests. Paper presented at American Educational Research Association, April, Chicago, IL. RAND, Washington, DC.

Linn, Robert L.
1975 Anchor test study: Final report; project report, and volumes 1 through 30. Anchor test study supplement, final report. *Journal of Educational Measurement* 12(3):201-214.
1993 Linking results of distinct assessments. *Applied Measurement in Education* 6(1):83-102.

Linn, Robert L., and V.L. Kiplinger
1995 Linking statewide tests to the National Assessment of Educational Progress: Stability of Results. *Applied Measurement in Education* 8(2):135-156.

Linn, R.L., L. Shepard, and E. Hartka
1992 The relative standing of states in the 1990 trial state assessment: The influence of choice of content, statistics, and subpopulation breakdowns in *Studies for the Evaluation of the National Assessment of Educational Progress Trial State Assessment*. Stanford, CA: National Academy of Education.

Loret, Peter. G., A. Seder, J.C. Bianchini, and C.A. Vale
1972 *A Description of the Anchor Test Study*. Princeton, NJ: Educational Testing Service.
1973 The Anchor Test Study: Administration of the Study. Educational Testing Service, Princeton, NJ. Available: ERIC microfiche collection, ED076672.

Los Angeles County Office of Education
1997 *Proceedings from the California Comparability Symposium, September 1997*. Los Angeles, CA: County Office of Education.

McDonnell, Lorraine
1997 *The Politics of State Testing: Implementing New Student Assessments*. CESE Technical Report 424. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

McLaughlin, Don
1998 Study of the Linkages of 1996 NAEP and State Mathematics Assessments in Four States. Final Report. John C. Flanagan Research Center Education Statistics Services Institute, American Institutes for Research. American Institutes for Research, Palo Alto, CA.

McLaughlin, Milbrey W., Lorrie A. Shepard, and Jennifer A. O'Day
1995 Improving Education Through Standards-Based Reform. Stanford, CA: National Academy of Education.

Messick, Samuel
1989 Validity. In *Educational Measurement*, Third edition, R.L. Linn, ed. New York: MacMillan Publishing Company.

Mislevy, R.J.
1992 *Linking Educational Assessments: Concepts, Issues, Methods, and Prospects*. Princeton, NJ: Educational Testing Service.

Mislevy, R.J., A.E. Beaton, B. Kaplan, and K.M. Sheehan
1992 Estimating population characteristics from sparse matrix sample of item responses. *Journal of Educational Measurement* 29:131-154.

National Center for Education Statistics
1998 *Linking the National Assessment of Educational Progress and the Third International Mathematics and Science Study: Eighth Grade Results*. Eugene Johnson and Adriane Siegendorf, NCES 98-500. Washington, DC: U.S. Department of Education.

National Research Council
1996 Evaluation of "Redesigning the National Assessment of Educational Progress." Committee on the Evaluation of NAEP, Commission on Behavioral and Social Sciences and Education, National Research Council, Washington, DC.
1997 *Education One and All: Students with Disabilities and Standards-Based Reform*. Lorraine M. McConnell, Margaret J. McLaughlin, and Patricia Morison, eds. Committee on Goals 2000 and the Inclusion of Students with Disabilities, Commission on Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

O'Neil, H.F., B. Sugrue, J. Abedi, E. Baker, and S. Golen
1992 *Final Report of Experimental Studies on Motivation and NAEP Test Performance*. Los Angeles: Center for Research on Evaluation, Standards, and Student Testing.

Pashley, Peter, Gary Phillips, et al.
1993 *Toward World-Class Standards: A Research Study Linking International and National Assessments*. Center for Educational Progress. Princeton, NJ: Educational Testing Service.

Peterson, Nancy S., Michael J. Kolen, H.D. Hoover, et al.
1989 Scaling, norming, and equating. Pp. 221-262 in *Educational Measurement*, Third edition, R.L. Linn, ed. New York: McMillan.

Roeber, Edward, Linda Bond, and David Braskamp
1997 *Annual Survey of State Student Assessment Programs. Fall 1996*. Washington, DC: Council of Chief State School Officers.

Rose, L.C., A.M. Gallup, and S.M. Elam
1997 The 29th annual Phi Delta Kappan/Gallup Poll of the public's attitude toward the public schools. *Phi Delta Kappan* 79(1):41-56.

Shavelson, R.J., G.P. Baxter, and J. Pine
1992 Performance assessments: Political rhetoric and measurement reality. *Educational Measurement* 21(4):22-27.
Shepard, L.A.
1994 The challenges of assessing young children appropriately. *Phi Delta Kappan* 76(3):206-210.
Shepard, L.A., R.J. Flexer, E.H. Hiebert, and S.F. Marion
1996 Effects of introducing classroom performance assessments on student learning. *Educational Measurement Issues and Practices* 15(3):7-18.
Shepard, L.A., and K.C. Dougherty
1991 Effects of High Stakes Testing on Instruction. Paper presented at the meeting of the American Educational Research Association/National Council on Measurement in Education, Chicago, IL. University of Colorado, Boulder.
Smith, Marshall S., David L. Stevenson, and Christine P. Li
1998 Voluntary national tests would improve education. *Education Leadership* (March):42-44.
U.S. Congress, U.S. Office of Technology Assessment
1992 *Testing in American Schools: Asking the Right Questions*. OTA-SET-519, February 1992. Washington, DC: U.S. Government Printing Office.
U.S. Department of Education
1997 *The NAEP Guide*. J. Calderone, L.M. King, and N. Harkay, eds. National Center for Education Statistics. NCES 97-990. Washington, DC: U.S. Department of Education.
1998 *Linking the NAEP and the TIMSS: Eighth Grade Results*. Eugene Johnson and Adriane Siegendorf. Project Officer, Gary Phillips, National Center for Education Statistics. NCES 98-500. Washington, DC: U.S. Department of Education.
Valencia, S.W., and K.H. Au
1997 Portfolios across educational contests: Issues of evaluation, teacher development, and system validity. *Educational Assessment* 4(1):1-35.
Waltman, Kristie K.
1997 Using performance standards to link statewide achievement results to NAEP. *Journal of Educational Measurement* 34(2):101-121.
Webb, N.M.
1995 Group collaboration in assessment: Multiple objectives, processes, and outcomes. *Educational Evaluation and Policy Analysis* 17:239-261.
Williams, Valerie, Kathleen Billeaud, Lori Davis, David Thissen, and Eleanor Sanford
1995 *Projecting to the NAEP Scale: Results from the North Carolina End-of-Grade Testing Program*. Technical Report #34. Chapel Hill, NC: National Institute of Statistical Sciences, University of North Carolina, Chapel Hill.
Wilson, Mark
1996 The California Comparability Study. In *Proceedings from the California Comparability Symposium, September 1997*. Los Angeles, CA: County Office of Education.
Wu, Grace, Mark Royal, and Don McLaughlin
1997 Development of a SASS 1993-94 School-Level Student Achievement Subfile: Using State Assessments and State NAEP. Feasibility Study. Education Statistics Services Institute, American Institutes for Research, Washington, DC.
Yen, Wendy
1996 Linking Tests for AB 265. In *Proceedings from the California Comparability Symposium, September 1997*. Los Angeles, CA: County Office of Education.
1998 Linking Assessment to NAEP and Providing Individual Student Scores. Paper presented to the Committee on Equivalency and Linkage of Educational Tests, March. CTB-McGraw Hill, Monterey, CA.
Yen, Wendy, and Steven Ferrara
1997 The Maryland School Performance Assessment Performance Program: Performance assessment with psychometric quality suitable for high stakes usage. *Journal of Educational and Psychological Measurement* 57(1):60-84.