Benchmarking and Alignment of Standards and Testing

CSE Technical Report 566

Robert Rothman, Jean B. Slattery, and Jennifer L. Vranek
Achieve, Inc.

Lauren B. Resnick
CRESST / University of Pittsburgh

May 2002

Center for the Study of Evaluation
National Center for Research on Evaluation,
Standards, and Student Testing
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

Project 3.1 Construct Validity: Understanding Cognitive Processes and Consequences
Lauren Resnick, Project Director, CRESST/University of Pittsburgh

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B960002-01, as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U.S. Department of Education.

BENCHMARKING AND ALIGNMENT OF STANDARDS AND TESTING

Robert Rothman, Jean B. Slattery, and Jennifer L. Vranek
Achieve, Inc.

Lauren B. Resnick
CRESST/Learning Research and Development Center, University of Pittsburgh

Abstract

The success of standards-based education systems depends on two elements: strong standards, and assessments that measure what the standards expect. States that have or adopt test-based accountability programs claim that their tests are *aligned* to their standards. But there has been, up to now, no independent methodology for checking alignment. This paper describes and illustrates such a methodology and reports results on a sample of state tests. In general, individual test items are reasonably well matched to the standards they are meant to assess. But the collection of items in a test tend to measure only a few of the less challenging standards and objectives. As a result, few state tests can be said to be well-aligned assessments of challenging standards.

Ten years ago, amid controversy about the wisdom of the idea and skepticism that it could become reality, diverse groups of Americans, committed to improving achievement in America's schools, promoted the idea of a standards-based system of education. The idea came simultaneously from multiple sources—a Presidential advisory panel, which included business and political leaders and the head of at least one of the major teachers' unions; an influential Commission on the Skills of the American Workforce, co-chaired by former secretaries of labor from both political parties; the National Education Goals Panel; and the National Council on Education Standards and Testing, a closely watched bipartisan panel led by two governors.

Three documents that lay out the concept of a standards-based system substantially influenced the national discussion over the next few years. One was a scholarly paper by Marshall S. Smith and Jennifer O'Day (Smith & O'Day, 1990) that described the major elements of such a system:

- standards (called, at the time, curriculum frameworks),
- school curricula,
- professional development,
- teacher education, and
- accountability assessment.

A second influential rationale for standards and the introduction of assessments geared to those standards came from the Commission on the Skills of the American Workforce (1990). Building on a benchmarking study of European and Asian education systems, this commission recommended (a) standards, (b) a high school certificate based on evidence of meeting those standards, and (c) commitment on the part of employers and higher education institutions to use the standards-based certificate as part of the basis for hiring and admission decisions. Although the Commission's recommendations concerned only the high school, its report became the basis for the New Standards Project, a consortium of states and urban school districts committed to developing a K-12 standards and assessment system based on state policies and district practices.

Much of the initial argument for the technical shape of a standards-and-assessment system came from a paper prepared by Lauren Resnick and Daniel Resnick for the National Commission on Testing and Public Policy (Resnick & Resnick, 1992). Based on a study of examination practices in European countries and their relationship to prescribed curricula, the Resnicks argued that examinations that created clear expectations for students and teachers both motivated and enabled more powerful teaching practices and created more equitable educational practices. The Resnicks also analyzed different forms of testing practices and argued for assessments based on direct observations of the kinds of performances desired from students (performance-based measures, projects, and portfolios—the "three P's"), rather than the indirect measures common in most American testing.

There were skeptics all around as debate grew over standards and standards-based systems. Many doubted that the country could—or should—move away from local control of curriculum content toward national, or even state, standards (national standards were an early consideration). The technical community raised doubts about the feasibility of performance assessment techniques and standards-referenced measurement, and many questioned whether the extra expense and

complexity was warranted even if such new forms of assessment were possible. In addition, advocates for minorities and educational equity feared that test-based credentials intended to raise overall achievement would, de facto, *de-credential* poor and minority children who could not meet the new criteria.

Despite these concerns, the "standards movement" took off in the succeeding decade. Today, little more than 10 years after the idea of standards entered the public arena, 49 of the 50 states have adopted statewide content standards for elementary and secondary schools. Forty-eight states have statewide tests or examinations at one or more grade levels, and more than half have set or plan to implement graduation or promotion requirements based at least partly on test results. In some cases, school districts—particularly large urban districts—have established standards, assessments, and promotion or graduation requirements on their own. Other such components of a standards-based system are on their way.

Given these statistics, we might be tempted to declare victory for the standards movement. But doing so would overlook both the rising public and professional "backlash" against some elements of the standards-and-assessment idea and the ways in which implementation has fallen short of the goals outlined a decade ago. Indeed these two—the backlash and the missing or weak elements of the standards-based system—are closely linked.

One source of backlash is the sense in many quarters that holding students accountable for meeting specific score expectations on tests in order to graduate from high school, or even to move to the next grade, is unfair. The sense of unfairness comes in part from the experience of middle-class families whose children have been performing well (or at least well enough) on traditional measures of achievement (including teacher grades and traditional standardized tests) and then suddenly score below announced cut points on newly introduced tests. In challenging the use of the new tests, many parents and other critics contend that the material on the tests does not reflect what students had been taught—or that if it does, it was because the curriculum was "dumbed down" to a low level. In either case, these critics charge that the use of the tests is unfair.

For poor, minority, or non-English-speaking students, the sense of unfairness is even more marked. Advocates for these students point out that these are the students most likely to fail the tests and thereby—in "hard" accountability systems—to be denied diplomas and other credentials. They are also the students

least likely to receive effective versions of the new, high-demand instruction that the new standards and tests call for. These students are, in other words, not receiving a fair *opportunity to learn* what they are being held accountable for. Formal standards for students' opportunity to learn were not included in either federal or state legislation establishing standards-based systems (over heated objections, at least at the federal level), but the idea of increasing students' opportunity to learn challenging content was an integral part of the original concept of standards-based education (National Council on Education Standards and Testing, 1992; Smith & O'Day, 1990). Indeed, advocates saw standards-based education as the critical lever for enhancing opportunities to learn and thus as a way of increasing equity in the educational system as a whole (Simmons & Resnick, 1993). Yet few, if any, states have put in place effective policies or resource systems for improving instructional quality (National Research Council, 1999).

One reason for the emergence of these two concerns—the concern that tests match curriculum only at a low level, if at all, and that low-income and minority students have not had the opportunity to learn the challenging content the tests demand—stems from the implementation of the standards-based systems. The theory behind standards-based education holds that standards should be rigorous and challenging, and that they should be specific enough to guide both teachers' and students' day-to-day work and the development of tests. In that way, standards become the lodestones for instruction, and tests are the measurement of the attainment of those lodestones. This calls for both high-quality standards and tests tightly *aligned* to those standards.

To what extent are today's standards and assessment systems meeting these quality criteria—and, indeed, how can judgments about system quality be made? Those are the questions addressed in this paper. We first describe in some detail a methodology for systematically comparing assessments to standards in order to determine the degree of alignment. The methodology was developed to provide guidance to states concerning the quality of their assessment systems. We then summarize the results of applying this methodology to a number of states and conclude by highlighting a set of issues that will face states and districts in a particularly sharp form as the new requirements of HR1 come into play.

The Problem of Alignment

Alignment is a widely used term (it occurs more than 100 times in the recently passed legislation reauthorizing the Elementary and Secondary Education Act) whose meaning appears simple, but whose technical definition has remained elusive. A dictionary definition of the term—being in agreement; a condition of close cooperation—captures quite accurately the ways in which alignment is commonly understood by teachers and policymakers. When applied to education, alignment refers to how well all policy elements in a system work together to guide instruction and, ultimately, student learning (Webb, 1997).

But how does one decide whether the elements in an education system, particularly those in the testing component, are working together? In the case of test-to-standards alignment, states—or other agencies commissioning a test—typically put a set of standards on the table and ask the test developers (sometimes internal to the state agency, sometimes external) to build a test that matches the standards. Not infrequently, test developers (or the contracting agency) begin the process by outlining a set of test specifications that call for certain numbers of items of certain types in order to sample the standards. Test developers typically respond by showing—usually in a matrix presentation of some kind—how items or tasks on the test match up to the standards statements. A matrix in which most cells are checked is offered as proof of alignment.

This seems pretty straightforward, but it masks myriad difficulties. For one thing, the terms used in standards documents are not completely constraining for test constructors. Test developers might claim that two items as different as the one shown in Figure 4 (presented later) and a 3-step arithmetic "story problem" each assess a standard calling for mathematical "problem solving." Secondly, standards are often written in terms so general that several different content elements are included in a single standard. The result is that different test constructors might focus their items on different parts of the standard. Finally, the cells in a matrix are typically checked if at least one item that assesses the standard appears on the test. A well-filled matrix, as a result, can mask a substantial imbalance in the extent to which different standards are assessed. At the request of several states interested in upgrading their standards and assessment systems, our group set out to develop a methodology that would go well beyond the checklist and matrix approach to

alignment.[1] We wanted to judge not only the quality of individual items, but also the overall qualities of the tests—the range, balance and degree of challenge represented by the set of test items as a whole. Furthermore, we sought a method that recognized that alignment is not an attribute of either standards or assessments per se, but rather of the relationship between them. And because it describes the match between standards and assessments, alignment can be legitimately improved by altering either one of them or both (Webb, 1997).

The Achieve Assessment-to-Standards alignment analysis is designed to answer the following questions:

- Does each assessment measure *only* content and skills reflected in the standards? In other words, can everything on the test be found in the state standards? This is a key element of fairness in a standards-based assessment system. Teachers and students need to be able to count on the standards as a guide to what they should focus on and know that they won't be assessed on something else.

- Does each assessment fairly and effectively sample the important knowledge and skills in the standards? In other words, to what extent does each assessment measure the key content and skills for that grade level? Tests that offer a disproportionate number of points for a small band of the content and skills specified in the standards, leaving the others out entirely or sampling them very lightly, are not well aligned.

- Overall, is each assessment sufficiently challenging? It is, for various reasons, easier to test the simplest knowledge and skills specified in standards—leaving complex concepts, extended reasoning and other higher level cognitive demands underrepresented in the test. Where the standards call for substantial high-level thinking, such low-demand tests are patently unaligned. Sometimes, however, a low-demand test can signal the need for more challenging standards.

Alignment is not a yes-or-no question; rather, it consists of a number of dimensions that collectively tell the story of the degree of match between the expectations states have for students' performance and the measure used to gauge whether students are meeting those expectations. The alignment analysis is a process of managing expert judgment. The process relies on experienced, knowledgeable educators, who bring their experience and knowledge to bear in

[1] Our work was conducted at the Learning Research and Development Center, University of Pittsburgh, and Achieve, Inc. Achieve, Inc., is an independent, bipartisan, nonprofit organization that helps states raise academic standards, measure performance against those standards, establish clear accountability for results, and strengthen public confidence in our education system.

applying the criteria for gauging alignment. The alignment protocol we describe here focuses and constrains the judgments of these experts so that their opinions can be aggregated with confidence.

The Alignment Protocol

The protocol reviewers use to analyze alignment considers four dimensions to be central in determining the degree of alignment between an assessment and standards.

Content Centrality

This criterion provides a deeper analysis of the match between the content of each test question and the content of the related standard by examining the degree or quality of the match. Reviewers assign each item to one of four categories based on the degree of alignment.

Performance Centrality

This criterion focuses on the degree of the match between the type of performance (cognitive demand) presented by each test item and the type of performance described by the related standard. Each item makes a certain type of cognitive demand on a student (e.g., the item requires a certain performance such as "select," "identify," "compare," or "analyze"). Reviewers assign each item to one of four categories based on the degree of alignment.

Challenge

This criterion is applied to a set of items to determine whether doing well on these items requires students to master challenging subject matter. Reviewers consider two factors in evaluating sets of test items against the challenge criterion: source of challenge and level of challenge.

Source of challenge attempts to uncover whether the individual test items in a set are difficult because of the knowledge and skills they target, or for other reasons not related to the subject matter, such as relying unfairly on students' background knowledge. Reviewers rate each item as having an appropriate or inappropriate source of challenge.

Level of challenge compares the emphasis of performance required by a *set* of items to the emphasis of performance described by the related standard. In addition to evaluating alignment, reviewers also judge whether the set of test items has a

span of difficulty appropriate for students at a given grade level based on the standards, the assessment, and supporting materials. Reviewers write a succinct summary of the level of challenge of each item set.

Balance and Range

Tests must cover the full range of standards with an appropriate balance of emphasis across the standards. No one assessment can measure the full range of knowledge and skills required by the state standards. Evaluating balance and range provides both qualitative and quantitative descriptive information about the choices states or test developers have made. *Balance* compares the emphasis of content supplied by an item *set* to the emphasis of content described by the standards. In addition to evaluating alignment, reviewers also judge whether the set of items emphasizes the more important content at the grade level. Reviewers write a succinct summary of the balance of each item set. *Range* is a measure of coverage or breadth (the numerical proportion of all content addressed).

Application of the Protocol

The process of applying the protocol is analogous to the scoring of performance assessments or portfolios. It begins with training in which participants go through each stage of the process using "anchor items" that illustrate each of the ratings. They then try to apply the criteria to a previously analyzed test, to measure their understanding of the criteria, before beginning with the "live" analysis.

In Achieve's application of the protocol, the raters selected to conduct the analysis represent a diversity of viewpoints including those of classroom teachers, curriculum specialists, and subject matter experts. As users or developers, some have had experience with large-scale assessments and many have had experience in standards development. The diversity of backgrounds is useful in judgments about the appropriateness of test items for particular grade levels, among other things. The participants meet in small groups to discuss their judgments, facilitated by an experienced group leader. The judgments are further structured by the tools developed for the process, including rubrics for judging the various dimensions of alignment.

As is typical in scoring of standards-referenced exams or portfolios, reviewers begin by becoming thoroughly familiar with the standards and tests they are to analyze. They read the standards document and supporting materials (e.g., sample

assessment questions, sample assignments, student work). They then take each test. Where items are open-ended, they note those aspects of a student response that might be elicited in the answer. While bringing their own understanding of reading, writing, and language skills to the task, reviewers attend to the constraints students face in taking the assessment (reviewers are aware of the time allotment and the tools and references that students may or may not use). Experience has shown this familiarization to be a critical step in the process. Reviewers internalize the scope and the structure of the test, as well as gain a preliminary, intuitive sense of individual items.

Following preparatory training, the protocol is applied in three stages: Stage 1, examining the match of the assessment to standards, item by item; Stage 2, examining the challenge posed by the test; and Stage 3, examining the balance and range of the test. Following those analyses, reviewers look at the test as a whole and then across grade levels, to make grounded judgments, based on the protocol data, about whether the state is measuring what it expects of students.

Stage 1: Item-by-Item Analysis

The item-by-item analysis has three steps: (a) a preliminary step in which the state's own alignment of items to standards is checked and where necessary, corrected, so that reviewers judging the items are always comparing them to the same standard; (b) a judgment of each item in terms of the *content centrality* for the standard to which it is matched; and (c) a judgment of each item in terms of *performance centrality* for the standard to which it is matched.

(a) Confirming the test blueprint. As a first step in examining the alignment of tests to standards, a senior reviewer matches test items to the blueprint the state or test contractor has prepared to guide the development of the test. The blueprint constitutes the state's or contractor's official statement of alignment. Confirmation of the test blueprint was originally thought to be a straightforward exercise meant to catch a few "discrepant" items that did not really map to the standards and which might confuse reviewers in the process of making judgments about content and performance centrality. The idea was to reassign these items before proceeding with the main set of judgments. However, in some instances, discrepant items were found to make up a substantial portion of the test. In a smaller number of instances, reviewers found items that did not match any standard. Such items were noted and eliminated from further analysis. To explain reassignments and elimination of items

to the states that had requested reviews of their tests, reviewers were required to provide a brief rationale for the reassignment or elimination of each item so treated.

In assigning items to standards, a *standard* was defined as the most specific level of outcome (i.e., the smallest "grain size") used by a state in delineating its content standards. For many states, these descriptors are called *objectives* and grouped under a more encompassing heading.

Where a standard is explicated in language or by example, the reviewer's job is to use that information to rate the assessment's match to the standard. Where the language of the standard is too broad or too vague for the reviewer to understand what kind of performance the standard actually expects, the reviewer's score or judgment will reflect the lack of clarity, as we will show in a moment.

The match between a test item and an objective does not depend on the difficulty of the content or the importance of that content. A match requires only that the question address the *same* content as the standard or objective. Mapping an item to the related objective is a yes-or-no decision.

To analyze the test blueprint, a senior reviewer

- examines each *item* in turn and identifies the match (if any) between the content of the test item and the content of one or more objectives; and

- checks his or her map against the developer's map and notes any discrepancies.

Because some states construct their test blueprints on the assumption that an item can measure several objectives, it was necessary to develop a decision rule regarding the handling of items that mapped to more than one standard. When an item is shown in the state blueprint to match more than one standard, reviewers evaluate one primary and one secondary match. When an item matches two or more objectives, the content of one of the objectives is usually most central to the item. Reviewers designate the most central match as the "primary match" and the next most central as a "secondary match."

Only primary matches are included in evaluating content and performance centrality and source of challenge. Secondary matches, as well as primary matches, are considered in evaluating level of challenge, balance, and range.

Figure 1 shows a representative example of the type of concerns that arise in analyzing test developers' blueprints or maps.

A poll is being taken at Baker Junior High School to determine whether to change the school mascot. Which of the following would be the best place to find a sample of students to interview that would be most representative of the entire student body?

A. An algebra class
B. The cafeteria
C. The guidance office
D. A French class
E. The faculty room

Source: NAEP 1996 public release item, eighth-grade mathematics.

This item was mapped as follows:

Data Analysis, Statistics and Probability
5. Use measures of central tendency, correlation, dispersion, and shapes of distributions to describe statistical relationships
b. Use the standard normal distribution

A reviewer might reasonably map this item, instead, as follows:

Data Analysis, Statistics and Probability
3. Understand and apply sampling, randomness, and bias in data collection
b. describe a procedure for selecting an unbiased sample

A reviewer might provide an explanation such as the following to justify his or her decision:

> The item requires students to think about the best situation available for obtaining a random sample. Each of the answer choices involves a certain kind of bias. The correct answer choice, B, is correct because it would produce the sample likely to have the fewest and least significant biases of the choices offered.

Figure 1. Test blueprint.

(b) Content centrality. After confirming the test blueprint, reviewers then examine each item and its relation to the standards. Judgments about content, or *content centrality*, go further than confirming the blueprint did in examining the content match between an item and a related standard. In making this analysis, reviewers consider both the specificity of the standard and the extent to which the content to be assessed is evident from a reading of the item (face validity).

Reviewers score each item for content centrality as follows:

2 Clearly consistent

1A Not specific enough (standard or objective is too broad to be assured of item's strong alignment)

1B Somewhat consistent (item assesses only part, and the less central part, of a compound objective)

0 Inconsistent

A "2" is awarded only when the standard is specific and the item clearly measures the content—for example, in language arts or mathematics—spelled out in the standard. Where the language of the standard does not specify the important language arts or mathematical content well enough to satisfy the reviewer that the content in the item is "fair game" for the assessment, the reviewer may not award a "2" for content centrality.

In some instances, the language of the standard is specific, yet the language arts or mathematical importance of the item itself does not fulfill the expectation set up by a reading of the standards. If the item shows a trivial use of the language arts content, or if the content of the item is clearly peripheral to the content in the targeted objective, the item scores a "0" for content centrality.

Reading selections are also evaluated for content centrality, to determine whether they reflect the type of reading expectations set forth in the standards. For example, if the standards call for reading both literary and informational texts, the selections must include both types of text in order to earn a "2" for content centrality.

Figure 2 shows examples of the kinds of issues that relate to determining the content centrality of individual items.

(c) Performance centrality. Judgments about *performance centrality* focus on the quality of the match between the type of cognitive demand presented by each item and the cognitive demand described by the standard the item is intended to measure.

Each item makes a certain type of cognitive demand on the student (e.g., "select," "identify," "compare," "analyze," "represent," "use"). A standard, if it is more than simply a list of topics, also specifies the type of performance required. Reviewers rate the match between the performance *actually called for* in the item and the performance indicated in the targeted objective. They judge whether the

Example 2a: Seventh-grade reading item

[Passage read is "When I Heard the Learn'd Astronomer" by Walt Whitman]

This poem is best classified as which of the following?

A. A sonnet
B. Epic poetry
C. Lyric poetry
D. A ballad

Suppose the above item has been mapped to the following standard:

> "The learner will analyze, synthesize, and organize information and discover related ideas, concepts, or generalizations."

Is knowledge of a literary type such as lyric poetry required by this item a central part of the standard? The standard does not specify the knowledge of literary forms so the judge is unable to say that this item is consistent with the standard. Mapped to the above standard, this item would receive a "1A" for content centrality. The standard is too broad to ensure strong alignment.

Suppose that the same item has been mapped to this standard:

> "Student can identify major literary forms."

The language of this standard is sufficiently specific for the reviewer to understand the item as being consistent with the content described in the standards. Because of this, and because the item measures important content implied by the standard, the item would receive a "2" for content centrality.

Example 2b: Eighth-grade reading item

Standard: Identify literary devices such as figurative language, allusion, sound devices, versification, foreshadowing, imagery, irony, and hyperbole, and determine the purpose of their use.

In this passage, the writer's tone is generally

A. sarcastic
B. hopeful
C. logical
D. sympathetic

In this example, the item questions the "tone" of the passage, a literary element not specifically referred to in the standard. One may consider, however, the author's creation of a specific tone as a literary device. Yet even acknowledging that knowledge of tone may fall under this standard, the item only assesses a part of the standard, that of identification with no determination of the purpose of the tone. For this reason, this item receives a "1B" for content centrality.

Figure 2. Content Centrality.

performance asked for in the item is what they would expect to see, having read the targeted objective. In practice, as they rate each item, reviewers usually compare the verb used in the item with the verb used in the targeted objective.

For instance, one part of a literature standard may be to "explain the effect of point of view." Explanation focuses on the student's ability to draw connections between the narrator and what he or she tells. Items that require students to do this are central to that literature standard and score a "2." Items in which the explanation is provided for the student score a "0" for performance centrality.

Reviewers score each item for performance centrality as follows:

2 Clearly consistent

1A Not specific enough (objective is too broad to be assured of item's strong alignment)

1B Somewhat consistent (the objective uses more than one verb, e.g., identify and analyze, but the item matches only one verb)

0 Inconsistent

An example of what is meant by weak performance centrality can be seen in Figure 2, example 2a. The performance called for in the standard is to *analyze, synthesize, and organize* information and *discover* related ideas, concepts, or generalizations. But the item asks students only to *identify* the form of the poem. None of the performances in the standard are demanded, or even permitted. In terms of performance centrality, therefore, the item would receive a "0."

Stage 2: Challenge

After rendering judgments about the match between individual items and the expectations laid out in the state standards, reviewers then turn to judgments about the challenge posed by the items. In this stage, reviewers consider two questions about challenge: (a) the *source* of the challenge posed by individual items, and (b) the *level* of challenge posed by sets of items.

(a) Source of challenge. Judgments about the type of challenge posed by test items and sets of items are intended to ensure that items are "fairly constructed" (i.e., the challenge stems from the subject matter and performance specified in the standard) and are not "trick" questions. Items with appropriate sources of challenge permit inferences that a student who does well on the item probably has a good grasp of the content targeted, and that a student who does poorly on the item

probably does not. Said another way, analyzing source of challenge helps to identify those questions likely to produce a "false positive" (i.e., a student may get the right answer for the wrong reason) or a "false negative" (i.e., the student gives the wrong answer, but in fact has the desired knowledge or skill). In an item with appropriate sources of challenge, the greatest challenges in the item lie in the content and type of performance targeted in the relevant objective.

Reviewers rate each item for source of challenge as follows:

1 Appropriate source(s) of challenge

0 Inappropriate source(s) of challenge

Items receive a "0" for source of challenge for two kinds of reasons. First, if an item scores a "0" for both content and performance centrality, then it is automatically scored "0" for source of challenge. This is because an item not matching the intent of the related standard in terms of the content or performance it is meant to assess is not a "fair" item—it is, in the end, measuring knowledge and skills that are not included in the standards for which students and teachers are accountable. Second, an item may be technically flawed. Common problems are having no answer or multiple correct answers; misleading graphics; incorrect labels on figures; reading items that are "text independent" (the student can answer the item correctly without having to read the passage on which the item is based); and mathematics items in which the reading is more a factor than the mathematics.

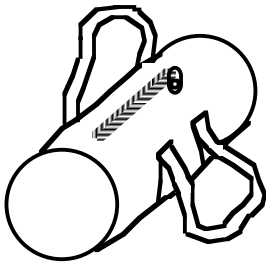
Figure 3 shows examples of items that introduce inappropriate sources of challenge. These examples show sources of interference that could be expected to reduce a reviewer's confidence that an item fairly taps a student's grasp of the content tested.

Example 3b points to the special attention that must be paid to the passages in a reading comprehension test. Source of challenge in a reading passage is derived from the level of comprehension demands placed on the reader. The most typical substitute for comprehension difficulty that test constructors use is a readability formula. Such formulas give a good estimate of vocabulary and grammatical complexity, but they do not take into account the cognitive load involved in processes such as making inferences, connecting information from different parts of a passage, or using background information. For this reason, the protocol asks reviewers to use their expertise and experience to make judgments about whether

Example 3a: Eighth-grade math item

How much fabric will you need to make a sports bag? Here is some important information about the sports bag.

- The bag is a cylinder made of these 3 pieces:
  the body, which is one rectangular piece of fabric
  two end pieces, each made from a circular piece of fabric.
- The bag is 60 cm long.
- The circular ends each have a diameter of 25 cm.



You must add an extra 2 cm all around each piece to allow the pieces to be stitched together.

Don't count the strap—it's made from different material.

What are the dimensions—the exact shapes and sizes—of the pieces that you will need to make one bag? Show your work step by step below.

Source: New Standards released item, eighth-grade mathematics.

This item is mapped as follows:

Geometry and Measurement Concepts

M2k: Models situations geometrically to formulate and solve problems.
M2a: Is familiar with cylinders.
M2d: Determines and understands length, area, and volume…computes areas of circles.

In this example, the challenges inherent in the context eclipse the mathematical challenge of the item. The context of sewing pieces of fabric together introduces the idea of "seam allowance." The difficulty of understanding and accounting for the "extra 2 cm" of seam allowance is considerable and is not part of the targeted mathematical idea: using understanding of a cylinder to model a situation geometrically.

This item would score a "0": inappropriate source of challenge.

Example 3b: Third-grade reading item

Standard: Makes inferences and draws conclusions.

This story tells you that many birds fly home in the springtime. Where have the birds spent the winter?

A. In warm places*
B. In sheltered trees
C. In cozy nests
D. In hibernation

In this case, the story implies that birds travel to warm places (*the correct answer), but it is never stated explicitly. For this reason, this item would be an adequate question for the standard. The source of challenge for this item, however, lies in the distracters. The student's understanding of the words "sheltered" and "hibernation" may present the item's difficulty, not the reading of the story. The item's source of challenge, therefore, lies in the vocabulary challenge, not the inference. Additionally, with adequate background knowledge, no reading of the text is necessary to determine the correct answer. For these reasons, this item receives a "0" for source of challenge.

Figure 3. Source of Challenge.

reading passages are appropriate for the grade level and whether they make unfair demands in terms of background knowledge.

To determine whether the source of challenge in a reading passage is appropriate, reviewers consider these questions:

1. Does the vocabulary seem appropriate for the grade level tested? Are words that may be typically judged to be outside of the reading vocabulary of the student accessible by using semantic, syntactic, or graphophonic cues?

2. Do the sentence structures seem appropriately simple or complex for the grade being tested?

3. If literary text is being presented, are the literary techniques accessible for the student at that grade? For example, if allusions are integral to the text, are students dependent on specific background information in order to comprehend the message?

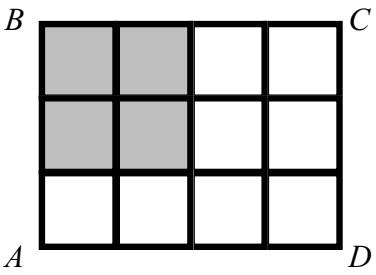
4. If a literary text, is the plot line sufficiently simple or complex for the grade level being assessed?

5. If the passage is informational, does the passage demand extensive prior background knowledge on the part of the reader in order to effectively comprehend the information?

6. Is the organizational structure of the informational passage likely to be familiar to the student; that is, is the structure representative of informational passages students are likely to be familiar with from reading texts, magazines, the newspaper, etc.?

(b) Level of challenge. The second judgment about challenge concerns the level of challenge posed by a set of items. In examining this question, reviewers make a global judgment about whether the collection of items in a test is appropriately challenging for students in a given grade level. Reviewers take into account the nature of the concepts assessed (some concepts are more readily understood than others) and the kind of thinking a student has to do to arrive at the answer (items may demand routine or concrete thinking as opposed to creative or abstract thinking). Figure 4 shows how apparently small changes in an item can substantially affect its level of challenge. An ideal item set would have items ranging from simple questions assessing accessible concepts to complex questions assessing concepts that are difficult to learn.

Example 4a: Eighth-grade mathematics item



In the figure above, what fraction of rectangle ABCD is shaded?

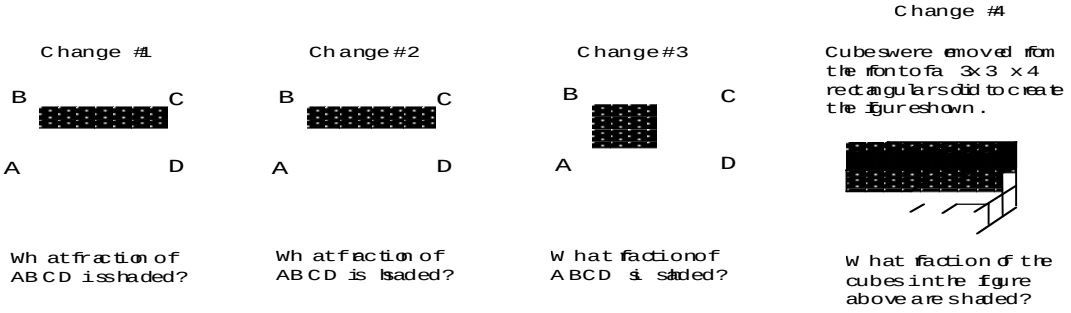
A. $\frac{1}{6}$; B. $\frac{1}{5}$; C. $\frac{1}{4}$; D. $\frac{1}{3}$; E. $\frac{1}{2}$

Source: NAEP 1996 public release item.

This item is mapped as follows:
Number Sense, Properties, and Operations

2. Represent numbers and operations in a variety of equivalent forms using models, diagrams, and symbols
   c. Use two- and three-dimensional region models to describe numbers.

Although the above item must be evaluated as part of a set of items for the challenge criterion, that evaluation starts with an examination of each item in the set. From this perspective, consider how the level of challenge in the above item might be significantly increased or decreased by any of four changes to the task that follow. (Note that all four changes result in items that would still meet the centrality criterion. However, each of the four of the changes would substantially affect the level of challenge of the item.)



Change #1 substantially reduces the challenge by showing only 3 regions for the figure.

Change #2 reduces the challenge by aligning the square units to make the shaded one-third easier to see.

Change #3 increases the challenge by requiring that the student make measurements to determine what portion of the figure is shaded.

Change #4 substantially increases the challenge by requiring that the student use a 2–D representation of a 3-D model, by requiring the student to use basic understanding of a rectangular solid and spatial visualization to calculate the number of cubic units in the figure, and by requiring the use of less commonplace fractions than in the original problem.

Figure 4. Level of Challenge.

The level of challenge judgment that reviewers are asked to make concerns the challenge of the item *set* compared with the expectations laid out in the standards. This definition differs from the notion of difficulty as traditionally inferred from scores on the items (*p*-values). The set of items and supporting material are examined in terms of whether most of the items are judged to be within an expected range of challenge, as delineated in the standard itself and any supporting materials. Where the language of the standard is too vague for the reviewer to understand the level of performance expected from students, this is noted in the evaluative remarks.

To determine whether the overall demand of an item set matches the expectations described in the standard or in supporting materials, reviewers are asked:

1. To what degree does the set of items make demands that are consonant with those specified in the standards?

2. Are the items skewed

    a. toward the more difficult *concepts* or the easier-to-grasp *concepts*?

    b. toward the most *cognitively demanding* or the least *cognitively demanding* of the objectives?

    c. where there are *compound* objectives that address concepts or cognitive demands?

    d. to the most demanding or least demanding part of the objectives?

For the set of items mapped to a given standard, reviewers provide a brief written evaluation of level of challenge.

Stage 3: Balance and Range

(a) Balance. Balance is a measure of how well a set of test items mapped to a standard reflects the emphases that the standard and its related objectives give to particular content. For this criterion, all items that map to a given standard—except those scoring "0" in content centrality—are considered as a set. The balance of items on a test should be representative of the balance of content and performance expressed in the standards. In other words, the relative importance that the test items give to content and skills should be the same as that stated in the standards. (As such, balance is distinct from "coverage.") For example, in a language arts test,

an ideal item set would have most items allocated to assessing the more important content and performances described by the standard and its related objectives.

To determine whether the overall balance of a set of items related to a standard matches the emphases in terms of content and performance described by the standard and related objectives, reviewers ask:

1. Which objectives in a standard seem to be over-assessed
    a. based on a reading of the standards (a straightforward issue of alignment)?
    b. based on what reviewers think is the most important content for the grade level (here reviewers bring their knowledge of the discipline to bear)?
    c. by items that are too much alike?
2. Which objectives in a standard seem to be under-assessed or unassessed
    a. based on a reading of the standards?
    b. based on what reviewers think is the most important content for the grade level?

Figure 5 provides some sample evaluations of balance for various standards. Reviewers' judgments form the basis for an overall evaluation of balance that is prepared by the team leader.

(b) Range. Range, in contrast to balance, is a measure of "coverage." For this criterion, all items that map to the standard—except those scoring "0" in content centrality—are considered as a set. Range, calculated arithmetically from the confirmed blueprint (Stage 1a), is the fraction of the total objectives mapped to a standard assessed by at least one item. Range values above .67 are considered good, whereas values between .50 and .66 are acceptable. A test may rate high on balance but low on range, or vice versa.

Some Results of the Analyses

We laid out two broad questions at the beginning of this report: (a) Is it possible to make nuanced, yet systematic judgments about the quality of assessments in relation to the standards they are meant to assess? (b) On the basis of a collection of such judgments, what can we say about the alignment and overall quality of today's test-based accountability systems?

Reading passages and items match the relevant standard
The set of questions presents a good map of the balance described by the relevant standards, in terms of those elements of the goals that can be fairly measured by an on-demand, multiple-choice test. The emphases evident in the range of questions were judged appropriate either to the nature of the selected reading passage or to the elements of greatest concern at this grade level (for example, an emphasis on "beyond text" rather than "intersentence" in Constructing Meaning).

Geometry items place a particularly heavy emphasis on vocabulary
Six of seven items include pictorial representations, none requires students to use models, none relies on manipulatives. Every geometry item ultimately hinges on knowing at least one special term. While this, in itself, is not bad, it is at variance with the standards document, which notes the use of manipulatives (twice), pictorial representations (twice), and models (three times), as well as vocabulary (twice).

Balance weak in "Probability and Statistics"
Four of seven "Probability and Statistics" objectives emphasize complex activities requiring the student to collect, organize, and display data, formulate questions and interpret information orally and in writing, display the same data in a variety of ways; discuss advantages and disadvantages of each form, and explore range, median, and mode as ways of describing a set of data. Much simpler processes are sampled by the items that mapped to these objectives: identify the question that cannot be answered by the data, identify one endpoint of a range given the other endpoint and the range itself.

Geometry items over represent a single shape and its component triangles
Items 14, 18, 23, and 27 are all based on squares decomposed into either two or four congruent right isosceles triangles. This means that 40% of all the conceptualization geometry items and about 33% of all the problem solving and applications geometry items are based on squares decomposed into congruent right isosceles triangles. Both a reading of the standards, and the professional judgment of reviewers, indicate that this is too much emphasis on one kind of set of shapes.

Assessment under represents number properties and measurement problem solving applications
Number properties and measurement problem solving applications are the two categories with standards that are focal points only at grades 4–6. The majority of these standards were not sampled in either the fourth grade or seventh grade assessment. Six of nine such number property objectives are not sampled. Two of three such measurement problem solving applications objectives are not sampled.

Figure 5. Balance.

We have described the Achieve alignment protocol in some detail in order to convey the complexity of the judgment task. We have also shown how it is possible to combine detailed, relatively low-inference ratings of specific items with more global expert judgments to yield evaluations of an assessment as a whole. The process we have described is quite expensive. It requires many hours of skilled judges' time. It can also be "politically expensive" because it yields evaluations that test developers or sponsors would rather not hear—since they have typically

claimed alignment for their tests, but the Achieve reports often tell them that alignment is very weak. Nevertheless, states continue to request evaluations of their assessment systems and are willing to pay for the service. They do this because they have come to trust the validity of the protocol and its application and find that they can defend the sometimes disappointing results to skeptical constituents and, indeed, use them as a basis for improving their assessment systems. The systematic technical analysis built into the Achieve protocol provides a grounded basis for judgments about the overall quality of assessments that makes even disappointing results useful.

As of the writing of this report, the Achieve alignment protocol has been applied to the English language arts (ELA) and math assessments in more than 10 states—about 20% of the nation. Each state has received a written report that describes the process and makes recommendations for how the state can address the issues raised in the report. The reports Achieve makes to states are confidential. For this reason we are not able to show here details of the analyses state by state. We can, however, report on some of the patterns of results. For this paper, we use results from five states whose data were made available for this report.

Item-by-Item Alignment

Figure 6 summarizes the scores on Content Centrality on elementary, middle, and high school tests in English language arts for these five states. Figure 7 provides the same summaries for math. Figures 8 and 9 summarize Performance Centrality scores for English and math, respectively.

Examining Figures 6 and 7, we see relatively high Content Centrality scores, except for ELA in state E. Figure 8 again shows state E as a low outlier on ELA Performance Centrality, with all the other states doing substantially better. Closer examination of state E's scores shows that for both Content and Performance Centrality, most of its ELA items received "1" (or intermediate) ratings. This is likely to happen when standards are stated in relatively global terms—and this was the case for state E. In other words, the problem for state E may not lie in the quality of test items, but rather in the quality of the standards—in this case, their global character. In math, state E joined the others with relatively high scores on both Content and Performance Centrality (Figure 9).

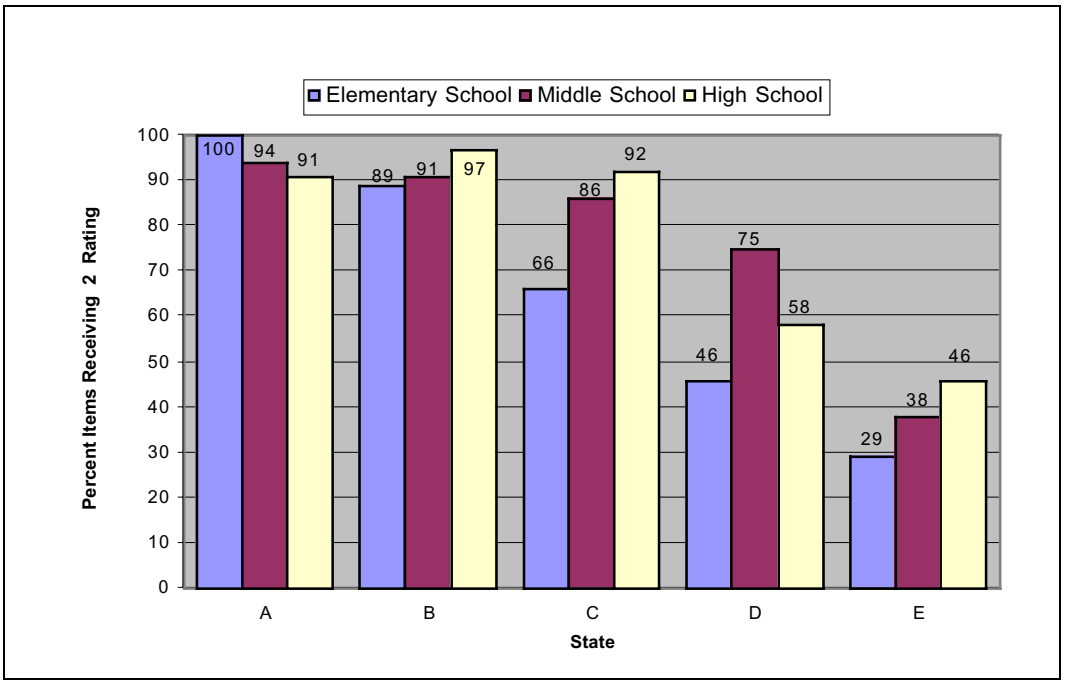


Figure 6. Distribution of Content Centrality Series-ELA.

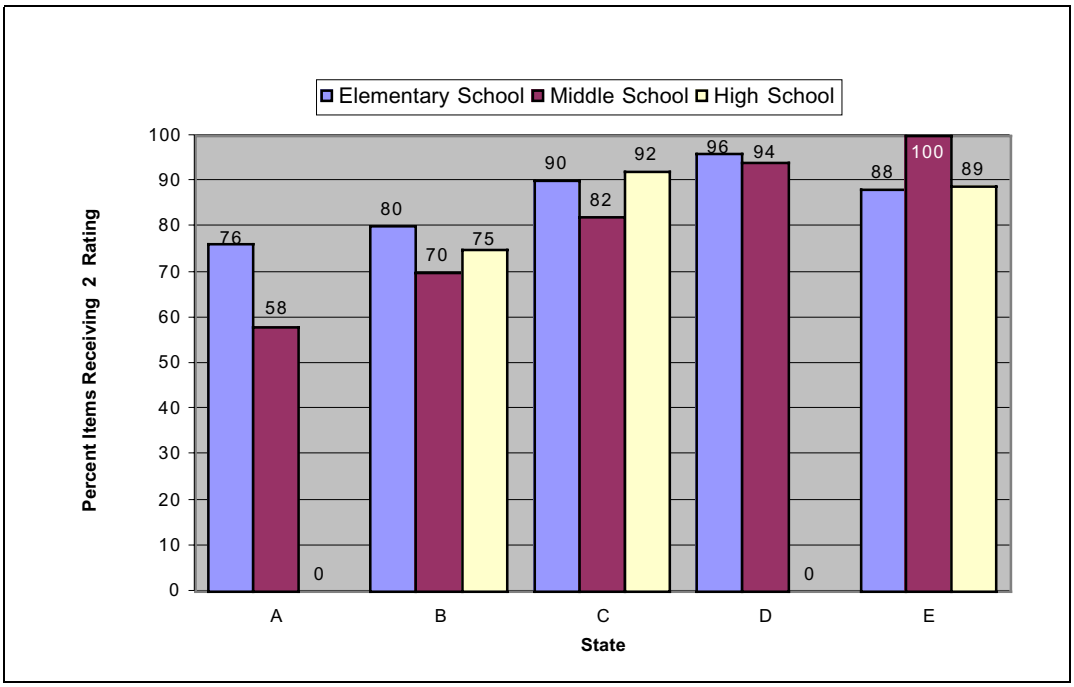


Figure 7. Distribution of Content Centrality Series-Math.

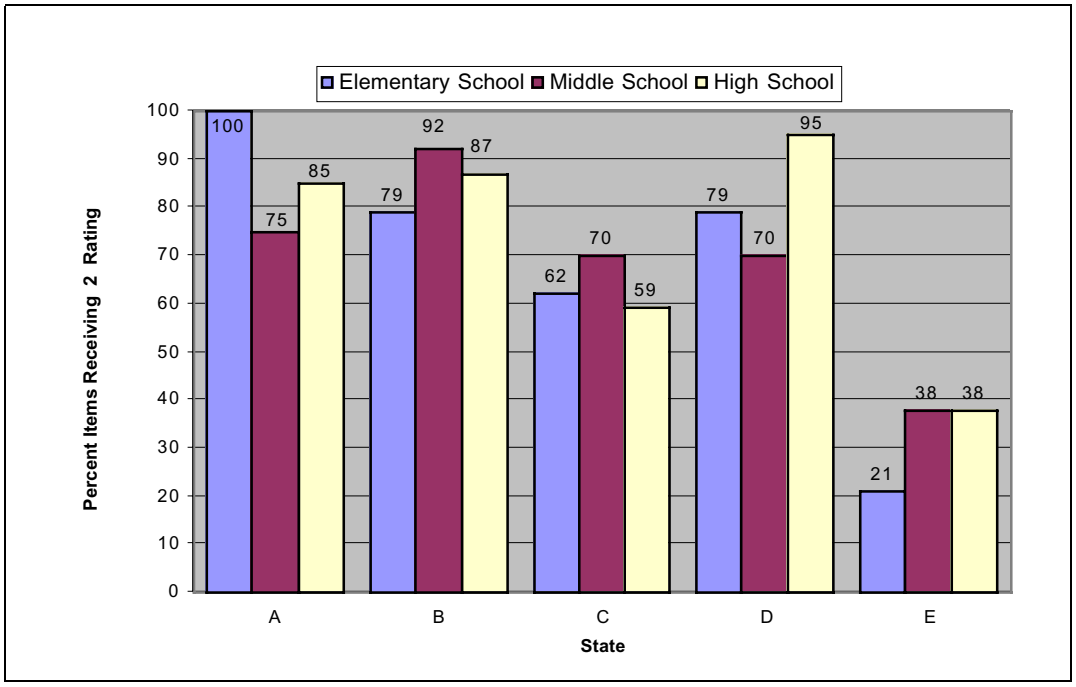


Figure 8. Distribution of Performance Centrality Series-ELA.

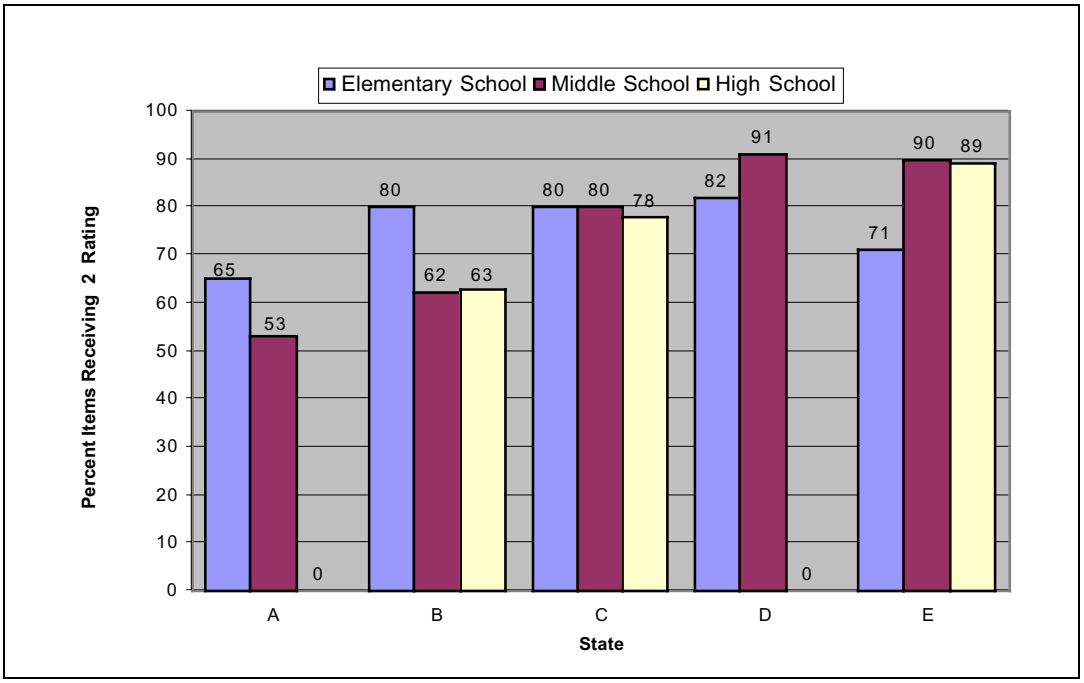


Figure 9. Distribution of Performance Centrality Series-Math.

Overall, then, the states that were analyzed proved relatively good at constructing or selecting items that matched the content and performances specified in their standards. A problem arises when standards are not adequately specific, however. We will return to this issue below.

Challenge

We turn next to the question of the appropriateness and degree of challenge in the tests. Figures 10 and 11 summarize the scores on individual items with respect to challenge. It is clear that, item by item, the tests do rather well in all states in both ELA and math.

What about the overall level of challenge of a test? For this we turn to the judgments that reviewers made about the challenge of the item *set* taken as a whole. These judgments were made standard by standard. If there were enough items for a standard to be judged on overall challenge, reviewers gave a rating of either *Appropriate* or *Too Low*. If there were too few or no items for the standard, reviewers noted that. With these aggregate judgments, we see a different picture of the quality of the state tests.

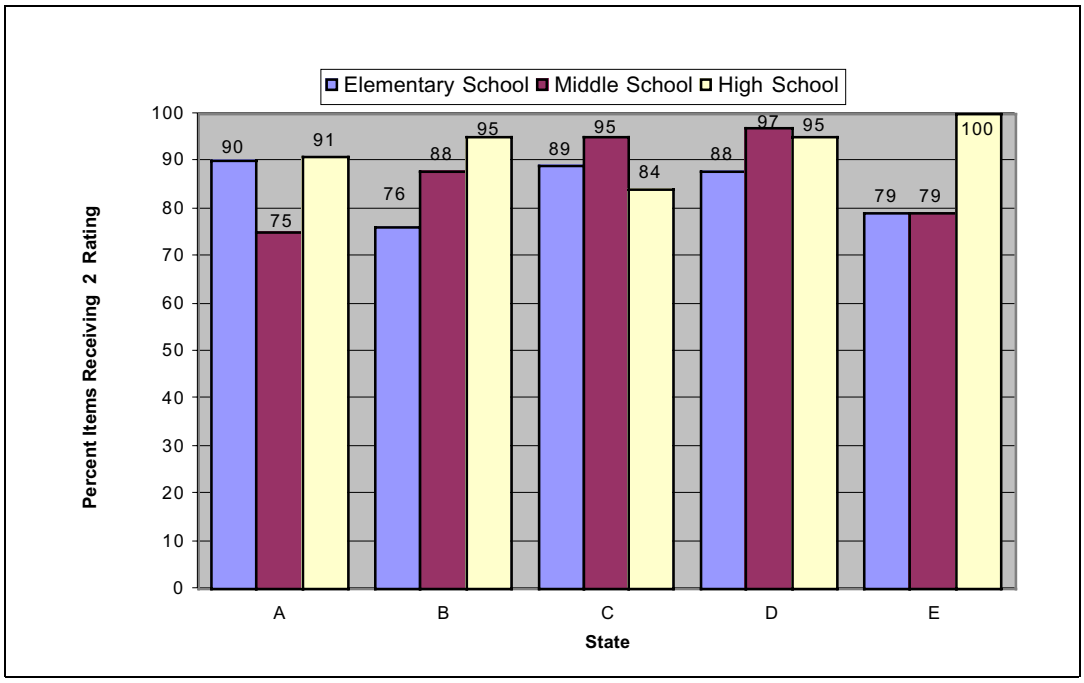


Figure 10. Distribution of Source of Challenge Series-ELA.

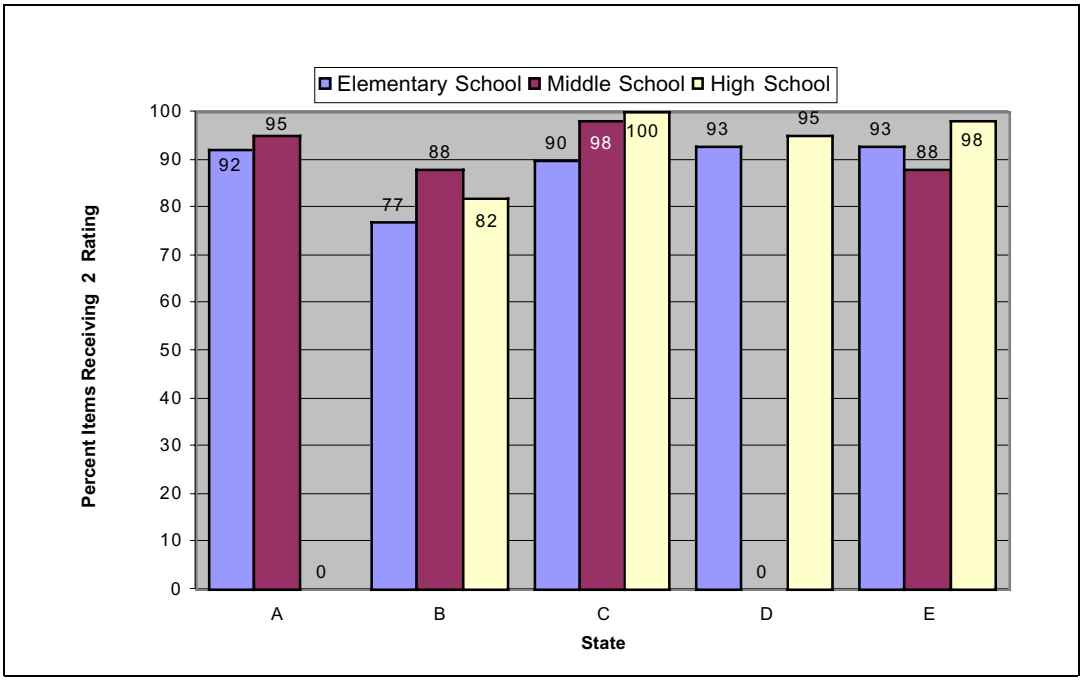


Figure 11. Distribution of Source of Challenge Series-Math.

Examples appear in Table 1. State A's scores for Content and Performance Centrality were good, as were its scores on Source of Challenge. Yet examination of the Level of Challenge judgments for state A's tests shows very few judgments of *appropriate* challenge—only 21% (7 of 33). Fifty-one percent of the standards were judged to be tested at too low a level of challenge, while 27% had too few items to allow a judgment.

State C's ELA-Reading tests provide a second example. The items on these tests received medium to high Content and Performance Centrality ratings and high Source of Challenge scores. But, as seen in Table 1, judgments for the tests' overall Level of Challenge were low. No standard was judged to have an appropriate level of challenge.

These examples are typical. Item qualities appear to be quite good, but with only a couple of exceptions, states are "tilting" their tests toward the least challenging of their own objectives.

Table 1

Example of Level of Challenge Judgments

Type of school	Appropriate	Low	None or too few	Total
		State A, Math		
Elementary	4	4	3	11
Middle	2	7	2	11
High	1	6	4	11
Total	7	17	9	33
		State C, English Language Arts-Reading		
Elementary	0	3	0	3
Middle	0	1	2	3
High	0	3	0	3
Total	0	7	2	9

Balance and Range

Balance and range are also judgments about tests as a whole. The Balance judgments, like those for Level of Challenge, are qualitative. Reviewers rate clusters of items mapped to a given standard for the degree to which, *as a set,* they constitute a balanced assessment of the standard. On the whole, the Balance judgments of our reviewers were close to their judgments for Level of Challenge. In other words, for most tests in most states, they found the tests not well balanced—with too much focus on some of the less important standards and objectives.

Range is a quantitative estimate of the proportion of objectives in a state's standards that are tested *at least once.* Table 2 shows the average Range scores for our five states in math and English language arts. These are quite varied. Most are at about the 50% range that we considered acceptable, but some are far below that. Since tests can only sample the range of objectives they are meant to measure, we would not expect to find 100% of objectives appearing on the tests. But the low numbers appearing in Table 2 raise some questions for testing policy. We know that some states—for example, California (which is *not* one of the states examined here)—have systematically chosen only a subset of their objectives for testing. If that was the case for the states examined here, the states did not tell us that. Had they

Table 2

Average Range Scores

Test and type of school	State A	State B	State C	State D	State E
ELA					
Elementary	.27	.75	.73	.52	.31
Middle	.28	.67	.58	.52	.36
High	.41	.67	.50	na	.31
Math					
Elementary	.52	.60	.53	.65	.75
Middle	.46	.62	.55	.75	1.00
High	.41	.56	.47	.73	1.00

done so, we would have modified our judging rules. Low scores on Range could also result from having too many micro-level objectives specified in standards documents. Schmidt, McKnight, and Raizen (1996) have referred to this as the "mile-wide, inch deep" problem in American math and science curriculum. Finding a workable middle path between overspecificity and the kind of underspecificity that we discussed above for state E will require careful attention in coming years.

Conclusion

This paper reports on a work that is very much in progress in the sense that states are actively working to develop and refine their standards and testing systems. Those efforts will probably increase substantially in the coming months and years as a result of the new federal legislation calling for annual testing of students. Standards, tests, and the relationship between them will be subject to increasing scrutiny, and we might expect a somewhat different evaluation of the overall status of alignment were we to begin our study now. Nevertheless, what we have reported here gives an important set of pointers to issues that will face states, and test contractors, over the coming period.

Taking our sample of states as a whole, and assuming they are representative of the nation, it is possible to conclude that states are working hard at aligning their tests to their standards. They have, for the most part, limited their tests to material that is in the standards—a primary requirement for a fair accountability test. Further—at least after our correction of the test blueprint—individual test items are generally quite well aligned to the standard or objective to which they are mapped.

This is shown in Content and Performance Centrality scores and in Source of Challenge scores.

But the good news ends here. With few exceptions, the collections of items that make up the tests that we examined do not do a good job of assessing the full range of standards and objectives that states have laid out for their students. What is included and excluded is systematic: The most challenging standards and objectives are the ones that are undersampled or omitted entirely. Standards and objectives that call for high-level reasoning are often omitted in favor of much simpler cognitive processes—low- or non-inference questions in reading, and routine calculations in math, for example. Thus, many of the tests in use by a state cannot be judged to be aligned with the state's standards—even though most of the items map to some standard or objective.

This particular pattern of non-alignment can have serious consequences for the kind of teaching that will occur in the states using such tests. As test-based accountability becomes more stringent, schools and teachers will match their curriculum and teaching ever more closely to what is on the tests rather than to what the standards say ought to count. The result will be an increasing focus on the low-demand aspects of the state's standards and a decreasing focus on the high-demand aspects that define a rigorous curriculum. To avoid this effect, states and test developers will need to attend carefully to questions of challenge and balance in their assessments.

References

Commission on the Skills of the American Workforce. (1990). *America's choice: High skills or low wages!* Rochester, NY: National Center on Education and the Economy.

National Council on Education Standards and Testing. (1992). *Raising standards for American education. A report to Congress, the Secretary of Education, the National Education Goals Panel, and the American people.* Washington, DC: U.S. Government Printing Office.

National Research Council. (1999). *Designing mathematics or science curriculum programs: A guide for using mathematics and science education standards*. Washington, DC: National Academy Press.

Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform in changing assessments. In B. K. Gifford & M. C. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement and instruction* (pp. 37-75). Boston: National Commission on Testing and Public Policy.

Schmidt, W. H., McKnight, C. C., & Raizen, S. A. (with Jakwerth, P. M., Valverde, G. A., Wolfe, R. G., Britton, E. D., Bianchi, L. J., & Houang, R. T.). (1996). *A splintered vision: An investigation of U.S. science and mathematics education. Executive summary*. Lansing: Michigan State University, U.S. National Research Center for the Third International Mathematics and Science Study.

Simmons, W., & L. Resnick. (1993). Assessment as the catalyst of school reform. *Educational Leadership, 50*(5), 11-15.

Smith, M. S., & O'Day, J. (1990). Systemic school reform. In M. E. Goetz & D. E. Mitchell (Eds.), *Politics of Education Association yearbook* (pp. 233-267). London: Taylor & Francis.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education* (Research Monograph No. 6). Madison: University of Wisconsin-Madison, National Institute for Science Education.