



State Collaborative on Assessment  
and Student Standards

# Combining Information from Multiple Measures of Student Achievement for School-Level Decision-Making:

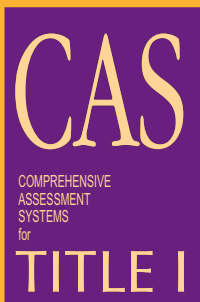
*An Overview of Issues and Approaches*

Phoebe C. Winter

*With the Study Group on Using Multiple Measures of Student Achievement*

Paul La Marca, Chair

November, 2001



*A Publication of the Council of Chief State School Officers*



The Council of Chief State School Officers (CCSSO) is a nationwide, nonprofit organization composed of the public officials who head departments of elementary and secondary education in states, the District of Columbia, the Department of Defense Education Activity, and five extra-state jurisdictions. CCSSO seeks its members' consensus on major education issues and expresses their views to civic and professional organizations, to federal agencies, to Congress, and to the public. Through its structure of standing committees and special task forces, the Council responds to a broad range of concerns about education and provides leadership on major education issues.

Because the Council represents each state's chief education administrator, it has access to the educational and governmental establishment in each state and to the national influence that accompanies this unique position. CCSSO forms coalitions with many other education organizations and is able to provide leadership for a variety of policy concerns that affect elementary and secondary education. Thus, CCSSO members are able to act cooperatively on matters vital to the education of America's young people.

The State Education Assessment Center is a permanent, central part of the Council of Chief State School Officers. The Center was established through a resolution by the membership of CCSSO in 1984. This report is part of a series sponsored by the Assessment Center's State Collaborative on Assessment and Student Standards (SCASS), Comprehensive Assessment Systems for IASA Title I. The series addresses issues related to the standards and assessment provisions of Title I.

Preparation of this report was supported in part by the U.S. Department of Education, Office of Elementary and Secondary Education. The views and opinions expressed in this report are not necessarily those of the Council of Chief State School Officers, the U.S. Department of Education, or the State Collaborative on Assessment and Student Standards.

## COUNCIL OF CHIEF STATE SCHOOL OFFICERS

G. Thomas Houlihan  
*Executive Director*



Wayne N. Martin  
*Director*

State Education Assessment Center



John F. Olson  
*Director of Assessments*

State Education Assessment Center



Phoebe C. Winter  
*Project Director*

State Collaborative on Assessment and Student Standards



Jan M. Sheinker  
*Coordinator*

Comprehensive Assessment Systems for IASA Title I

# **Combining Information from Multiple Measures of Student Achievement for School-Level Decision-Making: An Overview of Issues and Approaches**

**Phoebe C. Winter**

With the Study Group on Using Multiple Measures of Student Achievement  
**Paul La Marca, Chair**

State Collaborative on Assessment and Student Standards  
Comprehensive Assessment Systems for IASA Title I

Council of Chief State School Officers  
One Massachusetts Avenue, N.W.  
Washington, DC 20001

November, 2001

© Copyright 2001 by the Council of Chief State School Officers, Washington, D.C. This document may not be reproduced without the permission of the Council of Chief State School Officers, the copyright holder.

**State Collaborative on Assessment and Student Standards**  
**Comprehensive Assessment Systems for IASA Title I**  
**SERIES ON STANDARDS AND ASSESSMENTS**

*Opportunity and Change: An Overview of Title I of the Improving America's Schools Act of 1994.* Frank Philip, 1995.

*Designing Coordinated Assessment Systems for Title I of the Improving America's Schools Act.* Edward Roeber, 1995.

*Adequate Yearly Progress Provisions of Title I of the Improving America's Schools Act: Issues and Strategies.* Dale Carlson, 1996.

*Implementing the Adequate Yearly Progress Provisions of Title I in the Improving America's Schools Act of 1994.* Phoebe C. Winter, 1996.

*Using Existing Assessments for Measuring Student Achievement: Guidelines and State Resources.* Linda Hansche, Tom Stubits, and Phoebe C. Winter, 1997.

*Analyzing, Disaggregating, Reporting, and Interpreting Students' Achievement Test Results: A Guide to Practice for Title I and Beyond.* Richard Jaeger and Charlene Tucker, 1998.

*Handbook for the Development of Performance Standards: Meeting the Requirements of Title I.* Linda Hansche, with contributions from Ronald Hambleton, Craig Mills, Richard Jaeger, and Doris Redfield, 1999.

*State Standards and State Assessment Systems: A Guide to Alignment.* Paul M. La Marca, Doris Redfield, and Phoebe C. Winter, 2000.

# Preface

---

This report is the work of the State Collaborative on Assessment and Student Standards, Comprehensive Assessment Systems for IASA Title I (SCASS CAS) study group on using multiple measures of student achievement. It is intended to provide an introduction to issues involved in using the results of multiple assessments for school-level decision-making. We hope that it will serve as a resource to state and local educational agency administrators as they develop systems for school accountability that meet Title I requirements.

## STUDY GROUP ON USING MULTIPLE MEASURES OF STUDENT ACHIEVEMENT

Paul LaMarca (Chair)	Nevada Department of Education
Barbara Brandes	California Department of Education
Dale Carlson	Consultant, Council of Chief State School Officers
James Friedebach	Missouri Department of Education
Elaine Grainger	Vermont Department of Education
Barry Gribbons	College of the Canyons
James Grissom	California Department of Education
Paula Girouard	Massachusetts Department of Education
Ronald Houston	Delaware Department of Education
Scott Marion	Wyoming Department of Education
Doris Redfield	AEL, Inc.
Grace Ross	U.S. Department of Education
Alan Sheinker	CTB/McGraw-Hill
Jan Sheinker	SES Educational Services
Kathy St. Claire	Nevada Department of Education
Christine Steele	Wyoming Department of Education
Tom Stubits	Pennsylvania Department of Education
Carole White	Delaware Department of Education
Phoebe C. Winter	Council of Chief State School Officers



# Contents

---

INTRODUCTION	1
A Caveat .....	1
An Illustration .....	2
WHY USE MULTIPLE MEASURES OF ACHIEVEMENT?	5
Title I Requirements .....	5
Alignment and Technical Quality .....	6
Selecting Measures for Decision-Making .....	7
State A's Approach .....	7
COMBINING RESULTS	9
Three Approaches for Combining Information: Conjunctive, Compensatory, and Mixed .....	10
Combining Processes .....	11
KEY DECISION POINTS	13
An Illustration: State A .....	13
Computing Student Proficiency Levels .....	14
Computing School-Level Results .....	15
Summary .....	16
CONCLUSION	18
REFERENCES	19

---



Increasingly, state and local education agencies are building standards-based systems of education that focus on improving student learning in relation to well-defined academic content and performance standards. A critical component in this enterprise is measuring how well schools are educating their students to attain these standards. To accomplish this task, states are developing and implementing standards-based assessment systems, aligned with their academic content standards and designed to provide information about student performance in relation to pre-defined levels of proficiency such as “advanced,” “proficient,” and “partially proficient.” These systems contain instruments measuring achievement in a variety of content areas at a number of grade levels. A state might measure reading, language arts, mathematics, science, and social studies achievement at grades 4, 8, and 11, for example.

Title I of the Elementary and Secondary Education Act (ESEA) requires the use of multiple measures to judge the performance of schools relative to academic standards<sup>1</sup>. To judge how well schools are educating their students, the results of different tests<sup>2</sup> may be combined in some way to yield a single indicator or set of indicators about a school’s performance. For example, performance on two mathematics measures, a multiple-choice test and a portfolio containing work the student has done over the school year, might be combined to derive an overall indicator of student proficiency in mathematics. At the school level, student performance measures of science, mathematics, language arts, and social studies at grade 7 might be combined to yield an indicator of how well the school is educating its 7th grade students.

The purpose of this paper is to assist state- and district-level assessment, accountability, and Title I staff in making initial decisions about how to combine data for school-level decision-making. Its focus is on informed decision-making driven by measurement purposes and the uses of combined data. The paper describes critical issues to consider in designing, implementing, and using multiple measures and provides policy guidance for making decisions about how to combine the results of multiple measures of achievement.

## A Caveat

This paper does not deal with important technical issues; it focuses instead on the basic concepts and policy issues involved. It is the first report by the State Collaborative on Assessment and Student Standards, Comprehensive Assessment Systems for IASA Title I (SCASS CAS) on combining data to make decisions. The SCASS CAS will continue the work, illustrating various ways of combining data at the school level and the effects of using different types of models on several data sets. The group will explore complex technical issues that affect the results—for example, the number of measures used, the characteristics of the measures, implicit and explicit weighting schemes, and the relationships among the measures. The group also plans to examine the effects of including various non-cognitive indicators, such as attendance, in systems designed to describe school status and progress.

---

<sup>1</sup> At the time of publication, Congress was considering bills to reauthorize Title I of the Elementary and Secondary Education Act. The legislation under consideration does not reduce the requirements for multiple measures. If anything, the proposed legislation strengthens the need for states to carefully construct their systems and be cognizant of the educational, technical, and policy implications of their decisions.

<sup>2</sup> The words “test,” “measure,” and “instrument” are used interchangeably to mean any single measure of student achievement, regardless of format (e.g., open-ended items, multiple-choice items, portfolios, performance samples).

## An Illustration

State A has developed content and performance standards in reading, writing, and mathematics. The state administers measures in each of these areas in the spring of each year in grades 4, 7, 9, and 11. School districts are required to collect and report additional information in the three content areas, using criteria supplied by the state. The overall assessment system of state- and locally administered measures reflects the emphasis and depth of the standards in the three content areas. The 4th-grade assessment system is illustrated in Tables 1 and 2.

TABLE 1: STATE A'S 4<sup>TH</sup>-GRADE ASSESSMENT SYSTEM

Reading (5 Content Standards)	Writing (3 Content Standards)	Mathematics (6 Content Standards)
<u>Reading Measure A, Standards 2 to 5</u>  Test administered by state, scored on a scale of 100 to 300	<u>Writing Measure A, Standard 1</u>  Test administered by state, scored on a scale of 100 to 300	<u>Math Measure A, Standards 1 to 5</u>  Test administered by state, scored on a scale of 100 to 300
<u>Reading Measure B, Standards 1 and 4</u>  Locally administered portfolio of responses to reading selected books and articles, based on state-supplied criteria, scored on a scale of 1 to 6 for each standard, with a total score ranging from 2 to 12	<u>Writing Measure B, Standards 1 to 3</u>  Essay test, administered by state, scored on a scale of 1 to 4 for each standard, with a total score ranging from 3 to 12	<u>Math Measure B, Standards 1 to 6</u>  Locally administered portfolio of mathematics tasks, selected from state-supplied menu, scored on a scale of 1 to 6 for each standard, with a total score ranging from 6 to 36
<u>Reading Measure C, Standard 5</u>  Locally administered research project, based on specifications provided by state, scored holistically on a scale of 1 to 6	<u>Writing Measure C, Standards 1 to 3</u>  Locally administered portfolio of three "best" pieces from student's work over the year, selected based on state-supplied criteria, scored on a scale of 1 to 4 for each standard, with a total score ranging from 3 to 12	

TABLE 2: STANDARDS ASSESSED BY STATE A’S 4<sup>TH</sup>-GRADE ASSESSMENT SYSTEM

### Reading

Measure	Standards Assessed					Score Range
	1	2	3	4	5	
A. State Test		X	X	X	X	100 to 300
B. Portfolio	X			X		2 to 12
C. Project					X	1 to 6

### Writing

Measure	Standards Assessed			Score Range
	1	2	3	
A. State Test	X			100 to 300
B. Essay	X	X	X	3 to 12
C. Portfolio	X	X	X	3 to 12

### Mathematics

Measure	Standards Assessed						Score Range
	1	2	3	4	5	6	
A. State Test	X	X	X	X	X		100 to 300
B. Portfolio	X	X	X	X	X	X	6 to 36

Student scores on the locally administered measures are reported to the state, and the state uses these scores in combination with the state-administered measures to prepare score reports. Students receive a total score and a performance-level designation—Advanced, Proficient, Approaching Proficient, or Below Proficient—on each measure. Score reports at the school level show the total score distribution for each measure, the proportion of students scoring at each performance level, and a summary of student performance on each content standard within the three subject areas. The assessment results are then available for use in a variety of ways:

- Malcolm Kennedy, a 4th grade teacher in King Elementary School, has received the results of his students’ state and district assessments. He would like to compare the state results with his judgments of student learning.
- Carole Horton, the principal of King Elementary, wants to use the results to see how the 4th graders in the school are doing on the state content standards.
- Adrienne Falcon, State A’s accountability director, must recommend a method to the state school board for comparing year-to-year assessment results to see if schools are making adequate yearly progress in educating their students.

Each educator has specific questions in mind that the assessment results can help answer:

- Mr. Kennedy is asking, “What do the state assessments say about my students’ achievement? How well do my judgments of what my students have learned match the state results? Are there discrepancies? If so, how do I go about resolving them? What additional information can I collect that will

help me understand my students' achievement? What are the implications for my instructional practices?"

- Ms. Horton is asking, "What are my school's strengths and weaknesses on the state's reading, writing, and mathematics content standards? What are the implications of my 4th graders' performance for the school's curriculum? Are there areas I should follow up on by collecting additional information? What kinds of professional development should I offer to help my teachers target their instruction to enable students to meet the standards?"
- Dr. Falcon is asking, "How can we characterize a school's status and progress so that it can be compared from year to year? What criteria should we use to determine whether the school is making progress? Will the state assessment system alone provide the needed information?"

Mr. Kennedy, Ms. Horton, and Dr. Falcon will each need to combine the 4<sup>th</sup>-grade assessment results in different ways to fulfill their separate purposes for reviewing the results. They will have to evaluate their results to determine how well they answer their questions and decide whether they need additional information to fully answer their questions about student performance.

This paper focuses on Dr. Falcon's questions—that is, combining information (1) to characterize a school's current status and (2) to compare status from year to year to measure a school's progress. Because the results of assessments are used in different ways, we keep the uses illustrated by Mr. Kennedy's and Ms. Horton's questions in mind as we discuss school-level uses. The paper begins with the rationale for using multiple measures of student achievement and continues through decisions that are made along the way to producing school-level results.

# Why Use Multiple Measures of Achievement?

---

Multiple sources of information are used to improve the accuracy and utility of decisions all the time. Doctors diagnose illness and disease through multiple tests of different aspects of a patient's health. People buy cars using information about mileage, efficiency, size of the trunk, safety, style, price, and resale value. Teachers assign grades based on multiple classroom tests, homework assignments, projects, and observations of student behavior. In addition to providing information about achievement over time and about specific learning targets, multiple measures of achievement provide students with more than a single opportunity to show what they know and can do.

The *Standards for Educational and Psychological Testing* state that “a decision or characterization that will have a major impact on a student should not be based on a single test score” (AERA, APA, and NCME 1999, p. 146, Standard 13.7). This principle can be logically extended to making decisions that will have a major impact on schools, such as placing a school in program improvement or releasing a school from certain state regulations. Using multiple measures affords people the opportunity to make better decisions and more accurate inferences than they could if they looked at only one source of information. However, because designing and using an assessment system based on multiple measures to make decisions about schools is a relatively new endeavor, there is little guidance for how to go about it.

## Title I Requirements

While best measurement practice is delineated by the *Standards for Educational and Psychological Testing*, a major factor driving states to use multiple measures is Title I of the 1994 Elementary and Secondary Education Act. Title I requires that states measure student proficiency and the yearly progress of schools and local education agencies receiving Title I funds using the same assessment systems they use to measure the proficiency of all students in the state. Therefore, state assessment systems must meet the requirements specified by Title I legislation. They must

- measure and be aligned with the content and performance standards developed or adopted by the state;
- measure at least language arts and mathematics;
- meet established standards of technical quality;
- be administered at least once during each of three required grade spans (grades 3 to 5, 6 to 9, and 10 to 12);
- be administered to all students;
- provide appropriate accommodations for students with varied learning needs and language backgrounds;
- yield disaggregated group data;
- assess higher-order cognitive skills; and
- include multiple measures and approaches to measuring achievement.

Title I requires the use of multiple measures for several reasons: improving the alignment of the overall system of assessments to content standards, providing

information about the full range of student proficiency, improving the technical quality of results and the decisions made based on those results, and making the assessment system more accessible and fair than if only a single measure were used.<sup>3</sup>

### **Alignment and Technical Quality**

Multiple measures are often used to ensure that the assessment system is fully aligned with content and performance standards. Moreover, having more than one measure can contribute to the reliability, validity, and fairness of the decisions based on the results. States can use multiple measures to

1. increase the match of the assessment system with content and performance standards. The emphasis on high academic standards and higher-order thinking makes it unlikely that a single approach to assessment within a content area will adequately cover the knowledge and skills embodied in content and performance standards. Multiple approaches and measures can contribute to the degree to which the assessment system measures the range of content standards and the depth of content described in performance levels.<sup>4</sup>
2. increase the validity of student-level and school-level results. Alignment is a necessary condition for validity, at least in educational settings. For example, proficiency in science might be defined by the content standards as both knowledge of critical scientific facts and the ability to use the scientific process to test hypotheses and understand reports of scientific findings. A measure consisting only of questions about facts or questions about the scientific process would be a less valid measure of science proficiency than one that also required students to design an experiment and critique a newspaper article reporting on a new finding. The second set of measures, which is more aligned to the content standards, would allow the user to make a more valid inference about a student's standing in relation to the entire set of content standards.
3. increase the reliability (and validity) of student-level and school-level results. The use of multiple measures has the potential to increase reliability of results by increasing the number of items or tasks used to produce the results and by increasing the scope of content covered. When multiple instruments and formats are used, the items and tasks can include a greater variation of difficulty levels and elicit different types of responses. Allowing for multiple ways for students to demonstrate their knowledge and skills and allowing for a range of responses increases the likelihood of obtaining good measures of what each student knows and can do.
4. increase the fairness of assessment results. Assessment formats can vary when multiple measures are used, and the types of assessments employed (e.g., selected response, constructed response, or performance assessments) can more fully reflect different aspects of the content domain. The use of different types of formats enables students to demonstrate their full range of ability, a demonstration that could be constrained if only a single assessment format were used. By using multiple measures, states can increase the probability that the assessment items/tasks provide all students in the system the opportunity to demonstrate their proficiency in relation to content standards. A well-designed system of multiple measures can allow students who have learned the content in a variety of ways and mastered the content to varying degrees, students with disabilities, and students who are English language learners to demonstrate content knowledge and skill.

---

<sup>3</sup> For an in-depth discussion of the purposes for using multiple measures, see U.S. Department of Education, 1999.

<sup>4</sup> For a discussion of developing aligned assessment systems, see LaMarca, Redfield, and Winter, 2000.

5. increase the likelihood that schools will provide instruction in critical content areas and provide instruction in a variety of appropriate ways that emphasize skills reflected in content and performance standards. If one purpose of the assessment is to influence what content is covered in schools and how it is covered, then it is important that the measures represent the breadth and depth of the content. Research has shown that instruction is influenced by what is tested and how it is tested when assessment results are used to make important decisions (Resnick and Resnick, 1992; Shepard, 1991). If a single approach is used to measure student proficiency in a content area, it is likely that instruction will focus on that approach. Likewise, if a limited number of content standards are covered by the assessment system, it is likely that the tested content will be the focus of instruction.

### Selecting Measures for Decision-Making

Ideally, state staff and policymakers consider the various ways results of the state's assessment system will be used as they are designing the system. In reality, the uses and purposes of state assessment systems change over time. A system initially designed to provide student-level results might be used for school accountability. A system initially designed to provide school-level results may be augmented to provide information about individual student proficiency. Whether designing an assessment system from scratch or selecting measures to include in the system, "[t]he choice of the measurement instruments to be employed is the most important step in the entire process" (Ryan and Hess, 1999, p. 3; emphasis in the original).

A description of test development or selection procedures is beyond the scope of this paper.<sup>5</sup> However, the following critical questions should be considered when deciding what measures to use for making decisions about how schools are doing:

1. What are the primary purposes for collecting information about schools? What performance information is needed to fulfill the purposes?
2. How will the information be used? What decisions will be made based on school performance?
3. Do the measures adequately reflect the breadth and depth of the content standards? Do they provide information at each performance level? That is, are they aligned with the standards?
4. Do the measures provide reliable information about student performance?
5. Are there differences between the information needed from the assessment system and what the measures provide?
6. Does the system, as a whole, serve its intended purposes?

Some of the questions can be answered during the development or selection of the measures. Some can be answered only by trying out various techniques for combining data and analyzing the characteristics of the measures and how they contribute to decisions.

---

#### *State A's Approach*

---

State A revised its system to be more aligned with its content and performance standards. The state began with an assessment system that consisted of state-

---

<sup>5</sup> For criteria for reviewing existing instruments, see Hansche, Stubits, and Winter, 1997. For a summary of test development considerations, see Redfield, 2001.

administered tests in reading, writing, and mathematics. The tests included both multiple-choice and short-answer items, but they did not fully cover the state content standards. Instead, they covered the standards amenable to being measured by these formats. The state did have prototypes that districts could use to assess the remaining content standards. The state tests produced scale scores, but they were not referenced to performance standards. In making the transition to a more aligned assessment system, State A used its old system as a foundation and considered each of the critical questions posed above when evaluating its needs and modifying its assessment system.

**1. What are the primary purposes for collecting information about schools? What performance information is needed to fulfill the purposes?**

*State A's primary purpose is to judge how well schools are educating its students to meet state standards. To accomplish this, State A wanted to know how well its schools were educating students toward its reading, writing, and mathematics standards. The state selected four grade levels, 4, 7, 9, and 11, in which it would collect information about student performance.*

**2. How will the information be used? What decisions will be made based on school performance?**

*Assessment results will be used to determine whether a school is making adequate yearly progress. Results will also provide schools with information they can use to evaluate their curriculum and instruction. Student-level results will be reported to schools and parents to provide an indicator of how students are performing on state standards.*

**3. Do the measures adequately reflect the breadth and depth of the content standards? Do they provide information at each performance level? That is, are they aligned with the standards?**

*State A's original assessment system was not fully aligned with its content standards, and results were not reported according to performance levels. To improve the alignment of the system, the state created criteria for the development of local assessments, and it includes those measures in determining adequate yearly progress. To address the second major shortcoming, State A developed a system of performance standards.<sup>6</sup>*

**4. Do the measures provide reliable information about student performance?**

*State A instituted scoring procedures that maximize the consistency of scoring and evaluated the reliability of each measure and the reliability of the overall indicators used to judge school progress.*

**5. Are there differences between the information needed from the assessment system and what the measures provide?**

*State A made an initial determination that the alignment and rigor of the new system would provide the information it needed. The state planned a series of checkpoints when the system would be reviewed and could be modified as needed.*

**6. Does the system, as a whole, serve its intended purposes?**

*The state developed a plan for evaluating the validity of its proposed system for measuring adequate yearly progress. State A tried out a number of ways of combining its data, included some non-cognitive measures to see their effects, and selected the system that seemed to provide consistent, meaningful, and valid information.*

---

<sup>6</sup> For more information on developing performance standards, see Hansche, 1999.

# Combining Results

Multiple pieces of information about student proficiency can be combined at the student level to provide an overall picture of how the student is doing, and at the school level to provide an overall picture of how the school is doing. Results may be combined within a single content area, as is the case when results of two tests of science are combined into an indicator of overall science achievement, or across content areas, as is the case when science, reading, and mathematics results are combined into an indicator of overall academic performance.

Assessment purposes should drive the decision about whether to combine data from multiple measures. Table 3 provides examples of questions that might be the focus when combining results for a variety of purposes at different levels of aggregation.

TABLE 3: SAMPLE QUESTIONS BY BROAD PURPOSE AND LEVEL

	<i>Student</i>	<i>School</i>	<i>District</i>	<i>State</i>
<b>Accountability</b>	Has the student attained the standards?	Is the school making progress with all students, including students varying in income, ethnicity, gender, language proficiency, and disabilities status, thus attaining high standards?	Is the district or local education agency making progress toward the goal of all students attaining high standards?	Is the state making progress toward the goal of all students attaining high standards?
<b>Program Improvement</b>	What programs or services could the student benefit from?	What programs or services, including parent involvement, professional development, or extended day or year programs, need to be modified or added to enable the school's attainment of the goal?	What support does the district need to change or increase to help enable schools to attain standards?	What programs does the state need to add to meet the needs and enable attaining the goals?
<b>Instruction</b>	In what areas has the student done well, and in what areas does the student need more assistance?	What are the curricular and instruction problem areas?	In what specific areas does the district need to focus support for improving curriculum and instruction?	What areas of curriculum and instruction need additional attention at the state level?

Source: *Gribbons and Winter 1997.*

Although this paper concentrates on using information from multiple measures at the school level, most assessment systems collect individual student data and use that information to construct indicators of school performance. Therefore, this paper

includes discussions of techniques for combining information to produce student-level results as a means for obtaining school-level results.

### Three Approaches for Combining Information: Conjunctive, Compensatory, and Mixed

Three general approaches to combining data are used in educational settings: conjunctive, compensatory, and mixed.<sup>7</sup> A conjunctive approach requires satisfactory performance on each criterion (or measure) in order for overall performance to be deemed satisfactory; a compensatory approach allows less than satisfactory performance on some criteria to be offset by satisfactory or better than satisfactory performance on other criteria. A mixed approach combines the conjunctive and compensatory approaches by requiring a minimum level of performance on one or more measures and allowing performances above the minimum to compensate for each other. While each approach reflects a different underlying philosophy, practical and technical issues will often determine which approach is used.

To illustrate, let us return to State A's 4<sup>th</sup>-grade assessment system (see the section above entitled "An Illustration"). At the student level, reading proficiency is assessed using three measures. Each measure has cut-scores at each of the state's four proficiency levels: Advanced, Proficient, Approaching Proficient, and Below Proficient.<sup>8</sup>

- If a **conjunctive** approach is used, the lowest level earned on any single reading measure determines a student's overall proficiency level. A student must score at least at the Proficient level on each measure to be considered proficient in reading. If a student had a score pattern of Proficient, Approaching Proficient, and Proficient on the three measures, the student would be considered to have reached the Approaching Proficient level.
- If a **compensatory** approach is used, a student's overall proficiency level in reading is determined by some combination of performance on the three measures. Rules for allowing performance on one measure to compensate for performance on another would have to be developed through a standard-setting process. The "rule" might be as simple as converting all the reading scores to the same scale and taking the average score as the student's overall reading score. New cut-scores for each proficiency level would be referenced to the average score. Or the rule might be that if two out of three reading results are the same, that is the student's proficiency level; if the three results are all different, the middle result is the student's proficiency level. In this case, a student with a score pattern of Proficient, Approaching Proficient, and Proficient on three measures would be considered proficient in reading, as would a student with a score pattern of Proficient, Approaching Proficient, and Advanced.
- If a **mixed** approach is used, a student's overall reading proficiency level would be determined by some combination of performance on the three measures, with the condition that some minimum level of proficiency be achieved on each reading measure in order for a student to be considered proficient or above. For example, a mixed approach might include a rule that students earning a level of Below Proficient on any one of the reading measures could not be considered proficient, regardless of the level earned on the other two measures. With such an approach, a student with a score pattern of Below Proficient, Proficient, and

---

<sup>7</sup> Another method is the disjunctive approach, in which satisfactory performance on any single criterion is sufficient evidence of proficiency. This approach is not typically used in schools.

<sup>8</sup> The same three approaches can be applied to test scores rather than performance-level results. Performance levels are used here to illustrate the concept in a straightforward manner.

Advanced might be considered to be at the Approaching Proficient level (of course, such score patterns would suggest a need to review the measures for technical quality if such patterns were prevalent or to review the student’s work to understand the cause of the score disparity).

The approach makes a difference. Depending on the approach used, a student with scores of Below Proficient on Measure A, Proficient on Measure B, and Proficient on Measure C might have an overall reading proficiency level as shown below.<sup>9</sup>

Approach		Result
<b>Conjunctive:</b>	Lowest score determines proficiency level	Below Proficient
<b>Compensatory:</b>	Middle or most common score determines proficiency level	Proficient
<b>Mixed:</b>	Must score at least Approaching Proficient on all measures to be considered proficient	Approaching Proficient

These same approaches can be used to determine performance at the school level. For example, a conjunctive approach would require a school to meet minimum requirements in reading, writing, and mathematics; a compensatory approach could allow overall school proficiency in mathematics and writing to compensate for less-than-proficient performance in reading; and a mixed approach might require a minimum level of overall school proficiency in all three areas and allow for performance above the minimum in one area to compensate for lower performance in another.

Educators usually use compensatory and mixed approaches for judging the quality and degree of student learning. When student semester grades are based on their average performance over time, a compensatory approach is being used: good performance can offset bad performance. In some school districts, students can pass a course only if their classroom grades average to a passing grade and they also pass an end-of-course test; in this case, a mixed approach is being used, requiring minimum performance on two separate components (conjunctive), with grades being determined using a compensatory approach.

## Combining Processes

Scores from multiple assessments can be combined at either the student or the school level in two general ways.<sup>10</sup> The first is via a judgmental process. For example, experts in the content area review test scores or performance levels based on test scores and determine what combination of scores or levels is necessary for a student to fit into each category. The second is more heavily based on data such as the statistical relationships among the performance level results or the relationship between assessment results and an external criterion. Both judgmental and statistical processes can use either a compensatory, a conjunctive, or a mixed approach.

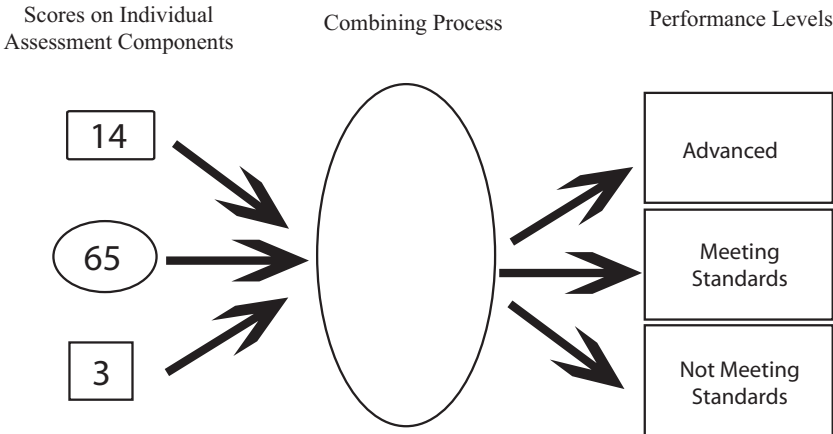
Student proficiency levels are typically estimated by combining student scores from individual measures. School performance may or may not be estimated by using the same combinatory approach or rules, but it is typically estimated after student-level performance is aggregated. Figures 1 and 2 illustrate the basic processes of combining

<sup>9</sup> These approaches are not constrained to determining proficiency in a single domain; they also can be applied across content areas such as math and reading.

<sup>10</sup> This is a bit of a simplification. No method is purely statistical, because judgment must enter into some stages of statistical processes, such as choosing a criterion or setting cut-scores on individual measures.

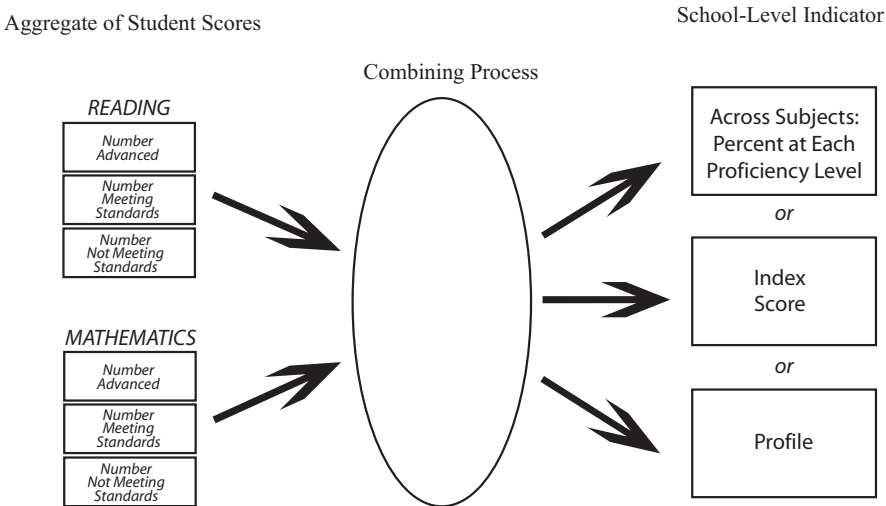
information from multiple measures at the student level and aggregating this information to judge school-level performance.

FIGURE 1: COMBINING DATA AT THE STUDENT LEVEL



At the student level, combining results of multiple assessments can be posed as a standard-setting issue. Haladyna and Hess (2000) evaluate compensatory and conjunctive approaches to setting standards (cut-scores) on assessments. Jaeger (1995) and his colleagues (Plake, Hambleton, and Jaeger, 1997) have developed judgmental methods for setting standards based on results of multiple tasks. Carlson (1996) has also proposed several judgmental methods for combining results from multiple assessments. Ryan and Hess (1999) described and compared several statistical methods for combining assessment results at the student level, again using simulated cut-scores on actual data. Once student data have been combined into performance levels, the next task is to aggregate and combine student-level performance into school-level results (see Figure 2).

FIGURE 2: COMBINING DATA AT THE SCHOOL LEVEL



## Key Decision Points

---

The ultimate dilemma facing those who need to combine data is arriving at a single decision, be it that a school is meeting expectations in helping all students attain challenging standards or that a student has demonstrated a particular level of proficiency. If 75 percent of students in a school are proficient or advanced in 4th-grade mathematics but only 48 percent are proficient or advanced in 4th-grade language arts, how should that school be categorized? If a student has demonstrated proficiency on two of three tests in mathematics but scored below proficiency on a third, what level of proficiency has the student attained? These two questions are related. While our focus is on the first question, the answer to the second question is essential to answering the first.

Our discussion is referenced to a specific purpose: using achievement data over time to determine whether schools are making progress from year to year in how well they educate their students. Although they stem from the same sources, the results that are used to make comparisons from year to year may or may not be in the same form as the results used to make decisions about how to modify school programs or policies. For example, combined reading and mathematics results, expressed as the percentage of students at or above the proficient level, may be compared from year to year to measure progress at the school, district, or state level. However, these results are likely to be relatively useless in determining a school's strengths and weaknesses for the purpose of improving how students are taught. Instead, schools might review student results in each content area, on each assessment, or on each standard to begin planning program improvement strategies. To be most useful, results describing school progress should be accompanied by more fine-tuned results or, at a minimum, schools should be given the information they need to break down the global results. At any rate, a clear definition of what the results of combined measures mean and how each particular set of results can be used is critical to promoting valid interpretations.

Initial policy decisions must be based on how the combined information will be used and the nature of the information to be combined. Once these policy decisions are made, a number of more practical decisions must be made about the details of combining information.<sup>11</sup> As always, decisions should be guided by the underlying purposes for assessment and accountability: to provide information that will promote and increase student learning.

### **An Illustration: State A**

To understand the types of practical decisions that must be made when combining information from multiple measures, we will return to State A's 4<sup>th</sup>-grade assessment system. Figure 3 recaps the way the state's system is configured.

---

<sup>11</sup> A caveat: Policy decisions must be made early in the process to guide later, more fine-grained decisions, but these policy decisions should be flexible enough that they can be modified if the data show that they are not working to meet the intended purposes of the assessment system.

FIGURE 3: ASSESSMENT SYSTEM CONTENT COVERAGE

	1	2	3	4	5
Measure A		X	X	X	X
Measure B	X			X	
Measure C					X

	1	2	3
Measure A	X		
Measure B	X	X	X
Measure C	X	X	X

	1	2	3	4	5	6
Measure A	X	X	X	X	X	
Measure B	X	X	X	X	X	X

### Computing Student Proficiency Levels

First, State A must have a method for obtaining an indicator of student proficiency within each content area as a whole (i.e., student proficiency level in mathematics, in reading, and in writing), by combining the results from each of the measures in each area. The following are two basic ways to determine a single, overall proficiency level using results from more than one measure:

1. Set cut-scores<sup>12</sup> for each performance level on each measure separately, then use the test-based performance levels to obtain a single proficiency level. For example, cut-scores on the mathematics assessments might be set as follows:

Measure A		Measure B	
Score	Proficiency Level	Score	Proficiency Level
100 to 125	Below Proficient	1 to 2	Below Proficient
126 to 181	Approaching Proficient	3	Approaching Proficient
182 to 246	Proficient	4 to 5	Proficient
247 to 300	Advanced	6	Advanced

In this case, if a student receives the same proficiency level on both measures (e.g., Approaching Proficient on Measure A and Approaching Proficient on Measure B), the student’s overall proficiency level is easy to determine. As part of the standard-setting process, judgments must be made about what overall proficiency level a student would receive if the two measures yielded different results (e.g., Approaching Proficient on Measure A and Proficient on Measure B), as illustrated below.

Proficiency Level on Measure B				
	Below Proficient	Approaching Proficient	Proficient	Advanced
Below Proficient	Below Proficient	?	?	?
Approaching Proficient	?	Approaching	?	?
Proficient	?	?	Proficient	?
Advanced	?	?	?	Advanced

<sup>12</sup> For a summary of procedures for setting cut-scores on individual measures, see Hambleton, 1999.

2. Use the scores on each measure in combination to set a standard resulting in a single proficiency level for each student (Carlson, 1996). For example, overall mathematics proficiency might be set as follows:

Score on Measure B	Score on Measure A							
	100 to 124	125 to 149	150 to 174	175 to 199	200 to 224	225 to 249	250 to 274	275 to 300
1	Below	Below	Below	Below	Below	Apprch.	Apprch.	Apprch.
2	Below	Below	Below	Apprch.	Apprch.	Apprch.	Apprch.	Apprch.
3	Below	Apprch.	Apprch.	Apprch.	Apprch.	Apprch.	Apprch.	Prof.
4	Below	Apprch.	Prof.	Prof.	Prof.	Prof.	Prof.	Prof.
5	Apprch.	Apprch.	Prof.	Prof.	Prof.	Adv.	Adv.	Adv.
6	Apprch.	Prof.	Prof.	Prof.	Adv.	Adv.	Adv.	Adv.

Within the state's system, establishment of student proficiency in reading and writing would be based on combining the three reading measures and the three writing measures, respectively, using similar techniques. Because three measures are involved in each content area, the level of complexity of the decision process increases. As the number of measures increases, the second technique described becomes more cumbersome. The two techniques illustrated above are based on some type of judgmental process. Ryan and Hess (1999) describe some techniques for combining scores based on statistical analyses.

Another approach to obtaining student proficiency levels in a content area is to combine information about individual content standards. In our illustration, reading standard 4 is measured by both the on-demand test and the portfolio; reading standard 5 is measured by the on-demand test and the research project. It may make sense to combine information at the individual content standard level first, and then obtain a proficiency level for each student based on content standard performance. Although this approach is not commonly used, it too could be employed based on professional judgments or through statistical techniques.

### *Computing School-Level Results*

Once each student has a proficiency level for each content area, student results must be aggregated to determine how well the school is doing. Typically, school results are based on the percentage of students who are in each proficiency level. In this case, school results would be based on first determining the number of students at each proficiency level within a content area and then combining these results across content areas, or by looking at each content area separately.

To illustrate, assume King Elementary School obtained the following results:

	Number (Percentage) of Students at Each Proficiency Level		
	Reading	Mathematics	Writing
Below Proficient	45 (18%)	35 (14%)	20 (8%)
Approaching Proficient	73 (29%)	53 (21%)	48 (19%)
Proficient	85 (34%)	75 (30%)	102 (41%)
Advanced	47 (19%)	87 (35%)	80 (32%)

1. If State A wanted to have a single description of how well the school was educating its students in reading, mathematics, and writing, a simple approach would be to calculate the overall proportion of students in each proficiency level across the three content areas:

	Reading	Mathematics	Writing	Total
Below Proficient	45 (18%)	35 (14%)	20 (8%)	<b>100 (13%)</b>
Approaching Proficient	73 (29%)	53 (21%)	48 (19%)	<b>174 (23%)</b>
Proficient	85 (34%)	75 (30%)	102 (41%)	<b>262 (35%)</b>
Advanced	47 (19%)	87 (35%)	80 (32%)	<b>214 (29%)</b>
<b>Total</b>	<b>250 (100%)</b>	<b>250 (100%)</b>	<b>250 (100%)</b>	<b>750 (100%)</b>

The total results would be compared with combined results from the previous year, and a decision rule about how much overall progress was required would be applied to determine whether the school was making satisfactory progress in educating its students.

2. If State A wished to look at each content area separately, results within each content area would be compared with the previous year's results. To reach a single judgment about whether the school was making satisfactory progress, decision rules would need to be made about the amount of progress required within each content area. The combining process comes into play in making these decision rules.

## Summary

State A's system illustrates, somewhat simplistically, the fundamental decisions for combining multiple measures, one at the student level and one at the school level. At the student level, the question is, "How will student results on multiple measures be combined to yield a single indicator of student proficiency in a content area?" At the school level, the question is similar: "How will overall student results in each content area be combined to make a single judgment about a school?"

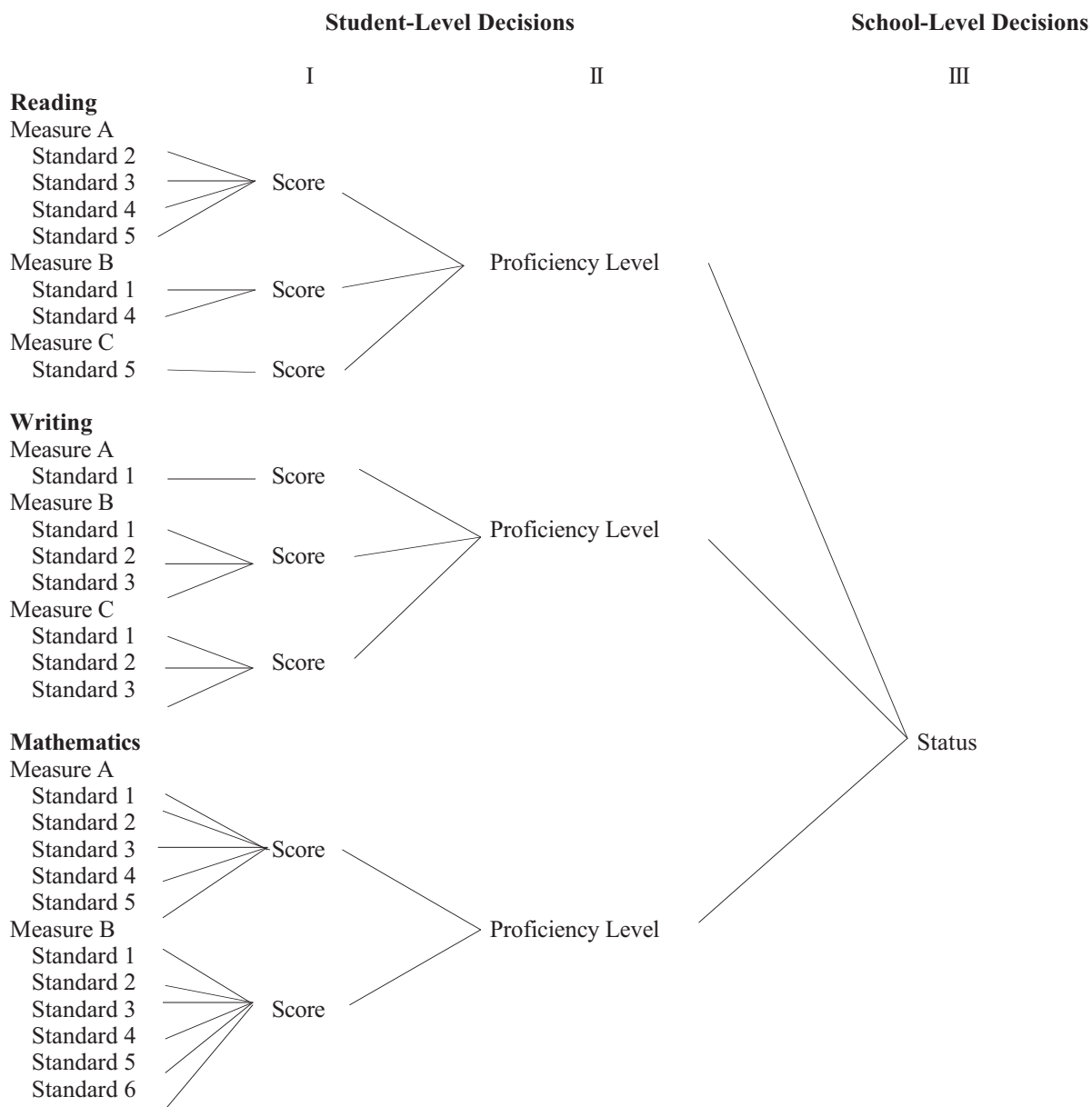
These decisions will have to be made based on the specifics of the assessment system, and they are as much a matter of judgment as statistics. For example, recall that the three reading measures in State A covered different, sometimes overlapping, content standards. Measure A addressed standards 2 through 5, Measure B addressed standards 1 and 4, and Measure C addressed standard 5. Depending on how the state's proficiency levels are defined, it may or may not make sense to obtain a proficiency level in reading on each measure. The second approach in the illustration, assigning proficiency levels based on the combination of scores on each measure, might be more appropriate.

The decisions that states must make in deciding to combine measures are summarized in Figure 4, using State A's system as an example:

- I: In developing each measure, the state had to decide how to measure each standard and how to combine the individual task scores into an overall score.
- II: The state had to determine how to combine the measures within each content area to obtain a measure of overall proficiency for each student in each content area.

III: The state had to decide how to use the combined/aggregated student results to characterize how well each school was educating its students.<sup>13</sup>

FIGURE 4: STATE A'S DECISION POINTS



<sup>13</sup> States may decide to add non-cognitive indicators, such as attendance or retention rates, to characterize how well schools are doing. The use of additional indicators is covered in later reports in the series.

## Conclusion

---

In this paper, we have discussed combining information at a particular point in time without paying much attention to how the information will be used in measuring school progress, although that is the ultimate purpose of the system. States will measure school progress by comparing results from one point in time with results from another point in time, but what comparisons the state wants to make should affect how the point-in-time result is obtained. In the illustration, reference is made to determining school progress from year to year in two ways: (1) comparing overall school results from year 1 to year 2 based on combined reading, writing, and mathematics results; (2) comparing results within each content area from year 1 to year 2. There are many other ways that states may look at progress from one time to another, and they will be discussed in companion pieces to this paper.

This paper has not dealt with important technical issues; it has focused instead on the basic concepts and policy issues involved. The SCASS CAS is sponsoring a project that will address some complex technical issues related to combining data from multiple sources. For example, how should scores from various measures be weighted? What are the effects of different weighting schemes? How does the reliability of the measures affect the combined result? Answering these kinds of questions will require the analysis of student data to see the effects of various decisions. The SCASS CAS will illustrate some methods for conducting such analyses in its future work.

## References

---

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- Carlson, D. (1996). *Adequate yearly progress in Title I of the Improving America's Schools Act: Issues and strategies*. Washington, DC: Council of Chief State School Officers.
- Gribbons, B., & Winter, P.C. (January 1999). Using Multiple Measures of Student Achievement. Unpublished proposal.
- Haladyna, T., & Hess, R. (in press, 2000). An evaluation of conjunctive and compensatory standard-setting strategies for test decisions. *Educational Assessment*.
- Hambleton, R.K. (1999). Setting performance standards on achievement tests: Meeting the requirements of Title I. In L. Hansche, *Handbook for the development of performance standards: Meeting the requirements of Title I*. Washington, DC: Council of Chief State School Officers..
- Hansche, L. (1999). *Handbook for the Development of Performance Standards: Meeting the Requirements of Title I. With contributions from Ronald Hambleton, Craig Mills, Richard Jaeger, & Doris Redfield*. Washington, DC: Council of Chief State School Officers.
- Hansche, L., Stubits, T., & Winter, P.C. (1997) *Using existing assessments for measuring student achievement: Guidelines and state resources*. Washington, DC: Council of Chief State School Officers.
- Jaeger, R.M. (1995). Setting performance standards through two-stage judgmental policy capturing. *Applied Measurement in Education*, 8, 15–40.
- La Marca, P.M., Redfield, D., & Winter, P.C. (2000). *State standards and state assessment systems: A guide to alignment*. Washington, DC: Council of Chief State School Officers.
- Plake, B.S., Hambleton, R.K., & Jaeger, R.M. (1997). A new standard-setting method for performance assessments: The dominant profile judgment method and some field-test results. *Educational and Psychological Measurement*, 57, 400–411.
- Redfield, D. (2001). *Critical Issues in Large-Scale Assessment*. Washington, DC: Council of Chief State School Officers.
- Resnick, L.B. & Resnick, D.P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B.R. Gifford & M.C. O'Conner (Eds.), *Changing Assessments: Alternative Views of Aptitude, Achievement and Instruction*. Boston: Kluwer Academic Publishers.
- Ryan, J.M., & Hess, R.K. (1999) *Issues, strategies, and procedures for combining data from multiple measures*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Shepard, L. (1991). Psychometricians' beliefs about learning. *Educational Researcher*, 20, 2–16.

U.S. Department of Education. (November 1999). *Peer reviewer guidance for evaluating evidence of final assessments under Title I of the Elementary and Secondary Education Act*. Washington, DC: Author.