Setting Coherent Performance Standards

State Collaborative on Assessment and Student Standards (SCASS)
Technical Issues in Large-Scale Assessment (TILSA)

Prepared by
Eric W. Crane
Phoebe C. Winter

THE COUNCIL OF CHIEF STATE SCHOOL OFFICERS

State Collaborative on Assessment and Student Standards

Technical Issues in Large-Scale Assessment

SETTING COHERENT PERFORMANCE STANDARDS

Prepared for the TILSA SCASS by

Eric W. Crane
Phoebe C. Winter

THE COUNCIL OF CHIEF STATE SCHOOL OFFICERS

THE COUNCIL OF CHIEF STATE SCHOOL OFFICERS

The Council of Chief State School Officers (CCSSO) is a nonpartisan, nationwide, nonprofit organization of public officials who head departments of elementary and secondary education in the states, the District of Columbia, the Department of Defense Education Activity, and five U.S. extra-state jurisdictions. CCSSO provides leadership, advocacy, and technical assistance on major educational issues. The Council seeks member consensus on major educational issues and expresses their views to civic and professional organizations, federal agencies, Congress, and the public.

Division of State Services and Technical Assistance

The Council's Division of State Services and Technical Assistance supports state education agencies in developing standards-based systems that enable all children to succeed. Initiatives of the division support improved methods for collecting, analyzing, and using information for decision making; development of assessment resources; creation of high quality professional preparation and development programs; emphasis on instruction suited for diverse learners; and the removal of barriers to academic success.

State Collaborative on Assessment and Student Standards

The State Collaborate on Assessment and Student Standards (SCASS) Project was created in 1991 to encourage and assist states in working collaboratively on assessment design and development for a variety of topics and subject areas. Division of State Services and Technical Assistance of the Council of Chief State School Officers is the organizer, facilitator, and administrator of the projects. SCASS projects accomplish a wide variety of tasks identified by each of the groups including examining the needs and issues surrounding the area(s) of focus, determining the products and goals of project, developing assessment materials and professional development materials on assessment, summarizing current research, analyzing best practice, examining technical issues, and/or providing guidance on federal legislation.

COUNCIL OF CHIEF STATE SCHOOL OFFICERS

Council of Chief State School Officers
One Massachusetts Avenue, NW, Suite 700
Washington, DC 20001-1431
Phone (202) 336-7000
Fax (202) 408-8072
www.ccsso.org

Acknowledgements

This report was written in collaboration with the TILSA Subcommittee on Coherent Achievement Standards, under the leadership of Liru Zhang, Delaware Department of Education.

Subcommittee on Coherent Achievement Standards

Liru Zhang	Delaware Department of Education
Jeff Barker	Georgia Department of Education
Selvin Chin-Chance	Hawaii Department of Education
Pamela Rogers	Kentucky Department of Education
William Insko*	Kentucky Department of Education
Fen Chou	Louisiana Department of Education
Joseph Martineau	Michigan Department of Education
Shungwon Ro	Minnesota Department of Education
Timothy Vansickle	Minnesota Department of Education
Walt Brown	Missouri Department of Education
Sharon Schattgen	Missouri Department of Education
David Abrams	New York Department of Education
Ronda Townsend	Oklahoma Department of Education
Ray Young	Pennsylvania Department of Education
Ellen Hedlund	Rhode Island Department of Education
Joseph Saunders	South Carolina Department of Education
Ed Miller	Texas Education Agency
Brenda West	West Virginia Department of Education
David Chayer	Data Recognition Corporation
Kevin Sweeney	Measured Progress

*Retired

The authors thank Huynh Huynh, Lou Fabrizio, Marianne Perie, and Carole Gallgher for their comments on an earlier version of this report, and Arthur Halbrook for his exceptional editing.

Table of Contents

Introduction

Provisions of the reauthorized Elementary and Secondary Education Act of 1965 (ESEA) hold schools accountable for implementing high standards in educating all students. Specifically, ESEA requires that students served by Title I funds be educated according to the same high standards as all students. The Improving America's Schools Act of 1994 (IASA), which amended the ESEA, required states to demonstrate that they have in place a coordinated system of content standards, performance standards,[1] and assessments that can evaluate student achievement on the content in reading and mathematics (Hansche, 1998). The assessments had to be administered in at least one grade in each of three grade ranges: 3–5, 6–9, and 10–12. Most states created state assessment programs or revised them to meet the IASA requirements through state-mandated, standards-referenced tests administered at three grade levels (Williams, Blank, Cavell, & Toye, 2005). The latest amendments to ESEA, as embodied in the No Child Left Behind Act of 2001 (NCLB), expand the role of state standards-based assessments in accountability. NCLB calls for testing in reading and mathematics at each of grades 3–8 and at one grade during high school.[2]

As states have begun developing academic content standards, performance standards, and assessments across contiguous grades, it has become clear that defining and promoting the systematic interrelationship of standards can strengthen the validity and defensibility of the assessment system (Cizek, 2005). Lissitz and Huynh (2003) describe a process for using *vertically moderated standards* as the basis for states' adequate yearly progress (AYP) decisions. In other recent articles, authors have explored this notion of standards' *coherence* (Lewis and Haug, 2005) or *consistency* (Mitzel, 2005). There is considerable overlap with how these authors have used these two terms. Lewis and Haug (2005) discuss coherence in terms of ''across-grade articulation of performance levels'' and expectations (p. 27). Lewis and Haug further explore coherence when looking at the impact of cut score decisions. The application of *coherence* to impact is problematic for focusing the term. In his use of *consistency,* Mitzel (2005) emphasizes results and impact somewhat more than do Lewis and Haug, but Mitzel's discussion of consistency includes performance levels and expectations.

Defining and promoting the systematic interrelationship of performance standards can strengthen the validity and defensibility of the assessment system.

We find it useful to distinguish the terms more clearly, following and expanding on the ideas of these authors. We will use *coherence* in reference to the system of performance standards and its components. Specifically, a system of performance standards is coherent when there are educationally sound relationships between performance levels and across grades. The term *consistency* is used to describe the impact of the cut scores and refers to a consistent (that is, close to the same) proportion of students at each performance level across grades. Coherence is thus the logical articulation of performance standards across grades; consistency deals with the quantitative results of impact

[1] Throughout this report, we use the term *performance standards* to refer to standards for student achievement or performance. The exceptions are direct quotes.
[2] NCLB also requires states to develop science content standards and assessments in each of three grade ranges: 3–5, 6–9, and 10–12.

data. While a coherent system can and often does lead to consistent impacts, coherence does not imply consistency. As Lewis and Haug (2005) report, a coherent system may include declining percents proficient across grades in a subject area. Certainly, any set of performance standards that lacks coherence invites problems:

> It will not be publicly acceptable, for example, for the 2004–05 fourth grade class to place 40 percent of its students at proficient or above and the same cohort to place only 20 percent of its students at proficient or above the next year. At the very least, this outcome would send a confusing message to parents and students. (Mitzel, 2005, p. 1)

Performance standards consist of three primary components:
1. ***performance levels***
2. ***cut scores***
3. ***performance level descriptors***

Performance standards consist of three primary components: (1) performance levels, which divide performance on an assessment into two or more qualitatively different categories (Title I of NCLB requires at least three), (2) cut scores, which define the score a student has to reach on a test to score within a performance level, and (3) performance level descriptors, or PLDs, which describe the knowledge and skills at each performance level on the assessment. PLDs are a critical tool for states to communicate and articulate performance standards.[3,4]

This report is intended to help states and interested readers navigate issues for developing a system of performance standards across contiguous grades in three ways: (1) by synthesizing important background information about setting coherent performance standards, (2) by summarizing existing work and promising practices in the states, and (3) by analyzing the issues concerning the development of performance level descriptors, including ways to maximize their effectiveness. A summary of key decision points for policymakers is in Appendix A, and a chart of selected state examples appears in Appendix B.

Background

For a look at ideas about coherence in standards, we begin with federal guidance for the development of standards and assessments under NCLB. The Standards and Assessment Peer Review Guidance of April 2004 includes important information about requirements for state assessment systems (U.S. Department of Education, 2004). The guidance document tells states about evidence they might submit in order to demonstrate they have met NCLB standards and assessment requirements and guides teams of peer reviewers who will examine the evidence and advise the U.S. Department of Education. The peer review guidance mentions *coherent* content standards and a *coherent* assessment system but does not directly mention coherence in terms of *performance* standards. However, the guidance states that performance

[3] Although some authors prefer the term, *achievement level descriptors,* we have found *performance level descriptors* to be more common in the literature, and we will use it throughout this report.

[4] This paper focuses on the issues surrounding the development of performance levels; issues such as how many levels there are and what they are called are typically matters of policy and will not be addressed here.

standards need to be articulated across grades, so coherence in performance standards is implied.

The guidance holds that coherent content standards ''must include only content that is meaningful with regard to the 'domain,' that is appropriate for the grade level specified, and that reflects clearly articulated progressions across grade levels'' (U.S. Department of Education, 2004, p. 8). Later in this paper, when describing procedures to build coherent performance standards, we apply variations of these three components.

Regarding the assessment system, the guidance maintains that state assessments must yield information that is coherent across grades and content areas. Although the reference to information that coheres across content areas is provocative, it is not explored further. Within a content area, the guidance gives an example similar to Mitzel's:

> For example, information gained from the reading/language arts assessment at grade 3 should be clearly and appropriately relevant to information gained from the reading/language arts assessment at grade 4 and subsequent grades. This does not require use of tests that are vertically scaled, but does imply the articulation of the standards from grade to grade. The content of the assessments and the achievement standards should be articulated across grades. (U.S. Department of Education, 2004, p. 24)[5]

Mitzel (2005) extends the discussion of consistency across subject areas. He argues that fundamental characteristics of human performance suggest that ''performance in one content domain should be similar to performance in another given a common group of students,'' but that ready guidelines for the degree of agreement do not exist (p. 10). Mitzel suggests that apparent disparities across subjects may be more acceptable to policymakers and the public than disparities across grades within a subject.[6]

Another issue that threatens the interpretation of test results reported based on performance standards is the variability across states in the meanings of the performance levels. If State A has 36% of its students scoring at proficient or above in math and State B has 72% of its students at proficient or above, there are several possible interpretations of the results. For example, State B's students may be doing better in math than State A's, or State A may have more rigorous standards than State B. Linn (2003) argues that such discrepancies seriously undermine the utility of using performance levels as a reporting mechanism. Linn points out that where cut scores are set is dependent on a variety of factors, including the method used, the panelists involved, and the context in which standards are set, maintaining that the resulting variability in percent proficient is ''so large that the term proficient becomes meaningless (p. 13).''

[5] While the peer review guidance mentions articulation across grades, it does not stress this component of standards and assessment systems in its review criteria.

[6] Mitzel uses the terms *horizontally* and *vertically* to refer to these respective situations.

State scores on the National Assessment of Educational Progress (NAEP) are often used to compare the relative rigor of state performance standards (see Ravitch, 2006, for a recent example). NCLB increased the frequency of state administrations of NAEP and mandated that states participate in the grades 4 and 8 assessments in reading and mathematics, making the role of NAEP more prominent. Although the validity of state to NAEP comparisons is often disputed, they capture the attention of policymakers and the press. Schafer (2005) gives an example of how performance data on state NAEP might be included in the context of developing vertically moderated standards.

To alleviate the lack of consistency in performance standards across states, Linn (2003) suggested that information about performance standards set in other states could be a useful component of the standard-setting process. However, how to introduce that information and how to set the context for the information is an area that needs some study. A comparison state may have a less stringent curriculum at a particular grade level, for example, which would make the interpretation of the state's percent proficient different than if the state used a curriculum of equal rigor. States vary in the content of their academic content standards, which could lead to valid differences in the interpretation of proficient and in the percentage reaching proficient across states.

A way to make performance standards more meaningful is to carefully define the knowledge and skills necessary to be classified in a performance level.

While Linn focuses on performance standards within grades compared across states, his observations apply to the use of PLDs to report scores within a state across grades. Linn (2003) notes, ''One of the purposes of introducing performance standards is to provide a means of reporting results in a way that is more meaningful than a scale score (p. 13).'' Carefully crafted PLDs can give meaning to student test performance in a way that a label alone (e.g., ''below proficient'' or ''minimally competent'') does not. Additional work is needed, however, to evaluate the validity of PLDs (e.g., a systematic study of the relationship between the PLDs, the assessments, and performance on the content domain both on tests and in the classroom) to strengthen their meaning and utility.

Conditions Contributing to Lack of Coherence

The absence of coherent performance standards poses a problem for states. Lack of coherence leads to a lack of interpretability and utility of test results in informing educators and other stakeholders. In this way, incoherent standards can lead to inconsistent results. If performance standards are not articulated well across grades, for example, instruction can suffer. Teachers and administrators miss a vital piece of information that can be used in designing curriculum and instruction that progresses smoothly through the grades. The lack of well-articulated performance standards across grades can result in teachers and parents having a distorted picture of how students are doing. For example, without a coherent system of standards, how can it be known whether the student whose performance drops from ''advanced'' to ''proficient'' genuinely lost ground? Could it be that the expectations were relatively higher in the upper grade? In short, if an ill-defined relationship exists between the

performance expectations at adjacent grades, then it may be difficult to make appropriate inferences about student progress.

Prior to NCLB, when testing in many states occurred at benchmark grades only, variation in the percentage of students at or above proficient was not necessarily cause for alarm.[7] One common example of cross-grade variation was a higher percentage of students in the proficient category at the earlier grades. Mitzel (2005) states that, in some cases, standard-setting panelists have been reluctant to brand younger students with a failing label; instead, panelists desire to identify only those students most in need of academic improvement. Setting higher standards at later grades also sends a message ''that instruction will need to improve for students to meet future standards'' (p. 2).

Furthermore, the standard-setting exercise often featured grade-level committees that stood apart from each other. It was common for the committees that were setting the cut scores for the state's tests at, say, grades 4, 7, and 10, to be convened only for general sessions such as at the welcoming portion of the agenda or perhaps at the farewell at the closing of the general session. The substantive work of setting standards happened in grade-specific groups, and, not surprisingly, the resultant performance standards frequently failed to cohere. With testing at every grade from 3 to 8, cross-grade variation and breaks in trend are more apparent and worrisome.

Consistent Results as Insufficient for Coherence

Consistent performance percentages do not by themselves imply coherent performance standards. States and standard setting panelists can set cut scores such that the results of any statewide academic tests can be put into consistent percentages. An extreme example makes this point. If tests at grades 3 through 8 all have, say, 56% of students scoring proficient or above, it is not helpful to call the standards coherent if we find out that the grade 3 test is in spelling, the grade 4 test in writing, the grade 5 test in mathematics, the grade 6 test in reading, the grade 7 test in science, and the grade 8 test in social studies! The standard setting must be predicated on more than smoothing numbers from impact data.

Content that is not clearly articulated across the grades is a threat to coherent standards within a subject area. Consistent performance percentages within a content area may suggest—but only suggest—that coherent standards are in place. In addition, if there are educationally sound reasons for the percentages reaching proficient to increase or decrease across grades (see discussion of Colorado's performance standards in the section, *Existing and Evolving State Practices*), then coherence does not even require consistent impact (i.e., percentage in performance category) across grades. Performance level descriptors that pinpoint expected cross-grade changes in performance are a necessary part of any system of coherent performance standards.

Content that is not clearly articulated across grades is a threat to coherent standards within a subject area.

[7] Cut scores that define other parts of the performance distribution (i.e., at or above basic, advanced) could also vary; however, under NCLB, their variation does not have the policy consequence of the cut score that defines proficiency. The role of other cut scores is discussed in the section, *Preliminary Cut Scores*.

In this section, we have discussed consistent performance percentages across grades and their place in a system of coherent performance standards. This may beg the question of whether it is possible to have an assessment system that detects when students in one grade are truly performing differently than students in other grades. With consistent standards, aligned curriculum and assessment, and a sufficiently large scale (e.g., a state or large school district), we would not expect to see a grade performing much differently than neighboring grades. In general, instructional efforts are assumed to be uniform across grades (Huynh & Schneider, 2005). However, a statewide initiative that brings differential resources to a particular grade, such as a reading initiative at grade 4, could cause such a phenomenon. Changes in test motivation that are systematic by grade, such as that postulated by NAEP for the grade 12 assessment, could be another factor that could lead to real differences in the test performance of an entire grade. In general, a fair, coherent system would ensure that students' performance is not a function of grade, but rather of their efforts and achievements (Lewis & Haug, 2005). It is our view that when state test results have an anomalous grade, it is much more likely to be due to inconsistencies in the standard setting than any other factor.

Importance of All Cut Scores

In light of NCLB's focus on having students reach proficiency, it is appropriate to emphasize coherence in the proficiency standard (Lewis, 2002). However, coherence throughout the range of performance must not be forgotten. The accountability systems in several states use an index approach with weightings assigned to performance in all categories, so crossing the basic or advanced threshold also matters. With recently announced federal flexibility for accountability models that are based upon growth, points along the performance continuum other than the proficient threshold may become more important in state accountability systems. An ideal set of performance standards for content that builds from grade to grade would have logical, well-defined relationships along the continuum from below proficient in grade 3 (or the earliest tested grade) through advanced in grade 8 and high school (or the latest tested grade). The gold standard for coherent performance standards requires all of the cut scores to be determined in thoughtful consideration of both the content and the impact data.

Existing and Evolving State Practices

Although most states[8] that have set performance standards on tests spanning contiguous grades have used a variation of the bookmark approach (Lewis, Mitzel, & Green, 1996), there are some differences in the details of states' approaches to developing coherent performance standards.[9] In several states, approaches are evolving. CCSSO's TILSA SCASS Setting Coherent Achievement Standards Subcommittee has been studying how states are developing coherent achievement

[8] Based on the 26 responses to a survey of all 50 state assessment directors, as of 3/10/06.

[9] Explanation of various standard-setting methods is beyond the scope of this paper. An excellent reference for the bookmark method and the various standard-setting designs is Cizek (2001).

standards.[10] The subcommittee has looked at published material from the states, as well as unpublished draft documents. A single driving question has motivated this exploration:

- Which state practices are most effective in developing coherent standards?

Three supporting questions serve to focus this initial question:

- How are states organizing their standard setting committees and procedures to support coherent standards?
- How are states using quantitative data to support coherent standards?
- How are states examining cross-grade differences in content and performance expectations to support coherent standards?

To help answer these questions, the authors asked all state assessment directors about the performance standards for their state's assessment system and the availability of reports on standard setting or the development of performance level descriptors. Our e-mail survey revealed that several states have published comprehensive technical material on their standard setting and PLD development work, and a few states have stated explicitly that coherence was a goal of their efforts (Delaware Department of Education, 2005; Huynh & Schneider, 2005; South Carolina Department of Education, 2005). Even in states that are not speaking about coherent standards explicitly, evolving practices are supporting coherence. In this section, we describe examples from 12 states. These examples include interesting lessons for states that have not yet completed the critical work of setting coherent standards.

Standard-Setting Committees and Procedures

Until recently, standard setting committees in most states have addressed material that is separated by multiple grades. As a result, the committees were grade-specific and did not come together to consider issues of cross-grade coherence. Even in states that have a history of contiguous testing, grade-specific standard setting committees have been the rule.

States have found that assembling panelists with expertise in adjacent grades supports coherence. Delaware assembled its reading and mathematics panels this way. Furthermore, in Delaware, participants in standard setting were encouraged to communicate with panelists at other grades during the standard setting (Delaware Department of Education, 2005). In South Carolina, the science cut scores for grades 3–6 were set by two panels that addressed grades 3–4 and grades 5–6, respectively (Buckendahl, Huynh, Siskind, & Saunders, 2005). Minnesota, Ohio, and Oklahoma also reported combining adjacent grades in building standard-setting panels. These examples are in keeping with Mitzel's (2005) advice:

Using standard-setting panels composed of members with expertise in adjacent grades supports coherence.

[10] The subcommittee's sponsorship and support of this paper are part of these initiatives.

> Compose panels with cross-grade membership. At each panel or table include participants from the grade level(s) above and below, where applicable. Consider, for example, having the same panel recommend standards for two or more grade levels at a time. Alternatively, consider asking panels to recommend standards at every other grade level, and use interpolation to set remaining targets. (p. 9)

Alaska's standard setting from May 2005 featured an innovative approach for composing and conducting a standard setting panel. The four-day standard setting for grades 3–9 kicked off with the entire group of panelists working together on grade 6. Following grade 6, approximately half of the group moved on to identify the cut scores for grade 5, while the other half worked on grade 7. After completing this work, the respective groups ''fanned out'' further, simultaneously working on grades 4 and 8 before completing the week's work by setting standards for grades 3 and 9 (Data Recognition Corporation, 2005a). The investment of time to establish a shared standard at grade 6 and then having the same panelists ''fan out'' to cover the grade 3–9 span is a novel approach to promoting cross-grade consistency.

There is moderate variation in the number of panelists that are included in standard setting. In general, the greater the number of independent panelists who are representative of the population of potential panelists, the more likely the results from the standard setting would generalize to the larger population of all potential panelists. For its 2005 standard setting, Ohio targeted 40 panelists per content area across all grades, and its panels included 42 panelists in reading and 46 in mathematics (Ohio Department of Education, 2005). Ohio involved more panelists than other states. Alaska set a goal of 24 panelists per content area across all grades. In Delaware, between 22 and 33 panelists came together to set standards in reading, mathematics, or writing in the standard settings that the state oversaw in July and August 2005. Pennsylvania's 2005 performance levels validation included 27 panelists in mathematics and 28 in reading (Data Recognition Corporation, 2005b).

States are taking other interesting steps regarding committee structure to meet policy aims and to increase the credibility and defensibility of their standard setting. While there is a long history of involving teachers, non-teacher educators, and non-educators in standard setting, having an entirely separate standard setting from the business community is an approach that few states are using. Florida had a non-educator panel conduct a separate science standard setting. The Florida panel included representatives from Pratt & Whitney, TECO Energy, Kennedy Space Center, Florida Aviation Aerospace Alliance, Postsecondary Faculty in Science, and Legislative Representatives. The recommendations from both panels were presented to the Florida State Board of Education at its February 2006 meeting.

Using Quantitative Data

Grade-by-grade standard setting procedures have routinely incorporated test data, either as part of the initial determination of cut scores, as in the use of item difficulty to present ordered items in the bookmark procedure, or as part of the consideration of impact of particular cut scores, as in the discussion of probable percentages of student at or above cut scores after initial recommendations have been made (Lewis, 2002). In attempting to develop performance standards that are coherent across grades, states have used data more often and in different ways. Item and test-level score data have been used to provide preliminary, or provisional, cut scores for panel review and to set boundaries for or otherwise constrain cut scores. Impact data have been used to show patterns of performance across grades and for setting standards through a form of *statistical moderation,* specifying a pattern of impact or performance distributions (most commonly, equal) across grades.

Preliminary Cut Scores

Presenting preliminary cut scores to panelists is a technique used by states that have established assessments and cut scores for some grade levels and are developing performance standards for tests in the other grades. For example, before 2002, Colorado had assessments in grades 4, 7, and 10 in writing and in grades 5, 8, and 10 in mathematics. When the state administered new tests in grades 3, 5, 6, and 9 for writing and in grades 6, 7, and 9 in mathematics in 2002, the state set cut scores on the new assessments and revised cut scores on the existing assessments (Lewis & Haug, 2005). Content experts reviewed patterns of performance across the five performance levels on the established tests and used this information to produce preliminary cut scores for all the tests. For writing, an equipercentile approach was used to set preliminary cut scores across grades. For mathematics, the content experts determined that a declining percent proficient as grade level increased was an appropriate model for Colorado. Standard-setting panelists were presented with the preliminary cut scores as part of a bookmark standard-setting procedure. Delaware, Pennsylvania, and Alaska also incorporated preliminary cut scores in their recent standard-setting or standard-validation activities (Data Recognition Corporation, 2005a, 2005b; Delaware Department of Education, 2005).

A technique some states have used is to present preliminary cut scores to standard-setting panelists.

For its testing program, Missouri set boundaries on acceptable cut score recommendations for proficient based on student performance on previous state assessments and on NAEP, to meet the requirements of state legislation (CTB/McGraw-Hill, 2005). The lowest percentage of Missouri students classified as proficient or above on NAEP grades 4 and 8 mathematics or reading assessments (26%) was used to define the highest acceptable cut score for proficient. The highest percentage of Missouri students classified as proficient or above on Missouri's existing grade 4 and 8 mathematics or reading assessments (44%) was used to define the lowest acceptable cut score for proficient. Thus, on each test, the cut score boundaries for proficient were set at the scores that would result in a maximum of 44% proficient or above and a minimum of 26% proficient or above.

States with vertically linked score scales are able to set constraints on how cut scores are set across grades. For example, Delaware used tables showing where proposed cut scores fell across grades on the vertical scale (Zhang, 2005). Panelists used this information to make their recommendations, increasing the probability that recommended cut scores would fall in a logical pattern across grades.

Using Impact Data

Another way Delaware used score data in its standard-setting procedure was to present patterns of student performance across grades to the panelists (Delaware Department of Education, 2005). During group discussion, panelists were asked to consider whether the impact data from preliminary cut scores made sense. Questions posed to the panelists for consideration were ‘‘(1) Do these observed patterns seem reasonable? (2) Do these trends represent what participants saw in the classrooms around the state? (3) Do their observations support this data?’’ (p. 10). Panelists were asked to present evidence about their positions regarding the reasonableness of the performance patterns, including any alternative patterns they thought were more reasonable.

Use of Trend Lines in Impact Data

For its 1999 state tests in English language arts and mathematics, South Carolina used student performance data to set cut scores for the grade levels between grades 3 and 8 (Huynh, Meyer, & Barton, 2000; Huynh, Barton, Meyer, Porchea, & Gallant, 2005). The state's test contractor conducted three rounds of the bookmark procedure to establish cut scores at grades 3 and 8. South Carolina examined the impact of its previous testing program and, based on performance patterns from those tests, determined that it was appropriate to use linear interpolation of the proportion of students in each level to set the cut scores at the intervening grades. For the 1999 tests, this process produced consistent results across grades 3 through 8.

(In)Stability of Cross-Grade Trend Data

An important consideration is the degree to which one can expect the relationships in performance patterns across grades to be stable across time.

An important consideration in using cross-grade impact data to set preliminary cut scores or as the sole basis of cut scores is the degree to which one can expect the relationships among grades to be stable across time. Cut scores that produce a specific impact pattern in one year (e.g., consistent percentages for each performance level across grades) may not produce that pattern in subsequent years. In South Carolina, the pattern of results has changed since the cut scores that produced consistent results were set in 1999, particularly in English language arts. In Colorado, which uses the same number of performance levels, the pattern has remained relatively stable since 2002. The two states have different test types: South Carolina's are custom-developed and Colorado's include norm-referenced test items as well custom-developed items. In addition, the pattern of initial impact of the cut scores differed in the two states. South Carolina's initial cut scores resulted in a larger proportion of students in the lowest level than Colorado's initial cut scores. For example, in mathematics, 47% of the students in grade 5

were classified at the lowest level in South Carolina while 12% of Colorado's fifth graders were at the lowest level. (Appendix C contains a summary of the test results for South Carolina and Colorado in the year in which they set cut scores and in 2005.)

These state examples indicate that the pattern of results evident today may or may not hold over time. At present, the assessment field lacks the data to know what contributes to stable performance patterns across years.[11] When impact data are used as the first or only consideration in setting cut scores, as when preliminary cut scores are presented to panelists or final cut scores are based on interpolation, the cut scores are based on assumptions about student proficiency. Preliminary cut scores set using an equipercentile model are based on the assumption that students in all grades are equally proficient at the time the cut scores are set. Preliminary cut scores set using interpolation are based on the assumption that student proficiency grows (or declines) across grades. The variation seen in the cross-grade patterns of student results across time throws these assumptions into question. Changes over time may be the result of instructional interventions targeted at specific grades, differential alignment between curriculum and assessment, or an otherwise unexplained cohort effect.[12] On the other hand, some of the changes in patterns of results may be due to characteristics of the test, such as equating error or differences in the test content over time.

An implication of instability in patterns of results across time is that the year chosen to provide the impact data for preliminary cut scores will have an effect on where the cut scores are set. For example, the test-independent conditions that cause change over time could also be in effect during the year used as a source of impact data, resulting in debatable assumptions being used as the basis for setting cut scores. One possible remedy is to have standard-setting panelists recommend cut scores before being shown cut scores based on equivalent (or increasing or declining) cross-grade impact. Discussion of the differences between content-based cut scores and data-based cut scores might bring out instructionally-based rationales for the variation, or it might reveal inappropriate expectations on the part of panelists. A second potential remedy is to review the long-term performance patterns of students on existing state tests before deciding whether to use impact data to set preliminary cut scores or for interpolation. If proportions of students in performance levels change steadily across time (e.g., in general, the proportion below proficient decreases steadily and the proportion proficient and above increases steadily, in all tested grades), it may be reasonable to use impact data to set preliminary cut scores. Lastly, states can consider educators' insights into cross-grade performance before setting preliminary cut scores, as Colorado did. All three strategies used together are likely to strengthen the validity of the assumptions about cross-grade student performance that underlie setting cut scores.

The year chosen to provide the impact data for preliminary cut scores will have an effect on where the cut scores are set.

[11] Stable performance *patterns* reflect consistent changes in percentages in levels across grades across time; this differs from stable performance *results*, which denotes no change in percentages in levels across time.

[12] When field test data are used as the basis for cut scores, lack of motivation and unfamiliarity with content arise as other factors that may threaten stability. We discourage states from relying heavily on impact data until the test is operational.

Although using test data as a central component of the standard setting process can promote the development of coherent standards, other components of the process can reduce the usefulness of the data. For example, in one state, cut scores recommended by the panels varied in their impact across grades by as much as 45 percentage points, despite the use of cross-grade panels and the sharing of impact data across grades. Group dynamics factors such as panel facilitation, panel composition, or the presence of a domineering panel member can affect the outcome of a standard setting meeting.

Analyzing Cross-Grade Differences in Content

Through statistical moderation, cut scores can be set to produce a smooth cross-grade trend in the percentage above the cut score. However, without any examination of the content, this process is solely a statistical exercise. Most respondents to our survey on standard setting practices identified the bookmark procedure as their state's standard setting method. The method does focus attention on the content, particularly the content whose difficulty is near a cut score. In its traditional application, the bookmark procedure does not address content across multiple grades.

The bookmark procedure, however, is readily adaptable to cross-grade analysis. If there are even a few items that have been tested at adjacent grades, the bookmark can be used to highlight changes in performance standards across grades. This procedure can be an efficient way to support coherent performance standards. In this adjusted method, an ordered item booklet (OIB) is assembled containing items from more than one grade, ordered by the item difficulties resulting from item response theory-based analysis. Bookmarks are placed separately to identify the cut scores for the grades represented in the expanded OIB. The multiple-grade or ''integrated'' OIB, as used in Mississippi, can be a powerful tool to highlight the changes in performance expectations across grades (Lewis, 2002).

Where a standard setting is revisiting grades for which cut scores already exist, stability is often a goal. The ordered item booklet can include a bookmark that has been placed at the current cut score. In Alaska and Delaware, this was the case, and panelists were given explicit instructions that any changes to the bookmark location must be justified by content considerations (Data Recognition Corporation, 2005a; Delaware Department of Education, 2005). Specifically, in order to justify a change to the bookmark location, the existing placement would need to be at odds with the grade level expectations. This requirement supports stability in the cut scores as well as coherent performance standards.

In Alaska and Delaware, panelists recommending changes to existing cut scores had to justify the change based on content considerations.

Analysis of Grade Level Expectations

A more systematic study of content was carried out in a pilot study in Delaware. Wise and his colleagues (2005) led an analysis of Delaware's grade level expectations (GLEs). Specifically, panelists focused on the differences between the GLEs at successive grades. Panelists rated the nature, importance, and clarity of differences. This process focused

attention on how the expectations at one grade relate to the expectations at the next grade. The results of the analysis informed the revising of Delaware's performance level descriptors.

In the Delaware study, panelists were asked to find, for each GLE at a given grade, the one or two most closely related GLEs from the previous grade. The panelists then identified the difference and used the *difference* as the subject of their rating. For each difference, they rated its nature (Broadened, Deepened, Same, or New), importance (Low, Medium, or High) and clarity (Not Clear, Minor, Clear). By focusing on the incremental changes in the GLEs, this process can hone in on the changes in performance expectations. The process even uncovered a number of GLEs (approximately 10%) where there was not a clear difference from the previous grade. These instances were flagged for the Delaware Department of Education review to see where further clarification of intended differences might be needed. In sum, this study generated critical information for the department as it was revising its GLEs and PLDs and moving more broadly to more coherent performance standards (Wise, Zhang, Winter, Taylor, & Becker, 2005).

The practices described in this section represent the efforts of several states that have been striving to build more coherent assessment systems. Along with the CCSSO's TILSA SCASS Setting Coherent Achievement Standards Subcommittee, we identified coherent performance level descriptors as an essential ingredient of a coherent assessment system and as a priority area for study. Attention will now turn to performance level descriptors and their role in the assessment system.

Developing Performance Level Descriptors

In the three-part system that also includes performance levels and cut scores, performance level descriptors (PLDs) have become a critically important tool for states to communicate performance expectations. PLDs are the link between a test score and content. The federal peer review guidance mandates them, but states have wide latitude as to their content and form. Nevertheless, descriptors that are not tied to grade-level content standards will not be acceptable to the U.S. Department of Education. In this section, we provide a framework for thinking about PLDs, including some examples of state practices.

PLDs have become a critically important tool for states to communicate performance expectations, linking the test score to academic content standards.

PLDs and Coherent Standards

Performance level descriptors can be a primary means by which states articulate coherent standards. Though NCLB does not specifically require that the *PLDs themselves* be coherent, the peer review guidance does require that *information from state assessments* be coherent across grades and content areas (U.S. Department of Education, 2004). Student performance, as classified into performance levels, is certainly a critical piece of information that assessments generate. Mills and Jaeger (1998) showed that the wording of PLDs could lead to substantially different cut scores and classification decisions. With the 1996 NAEP Grade 8

Science Assessment as a context, Mills and Jaeger used two sets of PLDs—those in place prior to testing and PLDs generated from examining a particular booklet or form from the assessment—as the bases for independent standard setting exercises. The same students' examinations were scored based on the two sets of standards, and the resulting classifications were strikingly different.

PLDs need to describe the competencies of each performance level in relation to grade-level content standards. Furthermore, they need to express a stable cross-grade progression within a single performance level. Some states maintain generic performance descriptors that refer to the student's readiness—or lack thereof—for the next grade level. By themselves, such generic descriptors are not sufficient to meet federal requirements. Generic PLDs may have value, however, as a precursor to more specific PLDs. Panelists who are writing PLDs may find useful the notion of ''ready for the next grade'' as they identify the specific competencies associated with a ''proficient'' performance level.

PLDs need to describe the competencies of each performance level, in relation to grade-level content standards, and express a stable cross-grade progression within a single performance level.

The National Assessment Governing Board (NAGB), which oversees NAEP, has adopted generic PLDs that it calls policy definitions, as well as grade- and subject-specific ones. The NAEP generic PLDs are shown in Table 1.[13]

Table 1. NAEP Policy Definitions (Generic Performance Level Descriptors)

Performance Level	Description
Basic	This level denotes partial mastery of prerequisite knowledge and skills that are fundamental for proficient work at each grade.
Proficient	This level represents solid academic performance for each grade assessed. Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real-world situations, and analytical skills appropriate to the subject matter.
Advanced	This level signifies superior performance.

Timing of Developing PLDs

There is some range of opinion as to the optimal timing of writing performance level descriptors. NAGB establishes performance level descriptors whenever a new main NAEP framework is adopted, and many state assessment systems have PLDs in place prior to standard setting. As the Mills and Jaeger study points out, there may be some misalignment between the PLDs and an actual form of the assessment. While the NAEP example is complicated by an unusually large number of forms, Mills and Jaeger's work suggests that there may be misalignment

[13] It should be noted that the NAEP performance levels are presented as ''developmental in nature and continue to be used on a trial basis'' (U.S. Department of Education, National Center for Education Statistics, 2005, p. 231). There is not universal agreement on what constitutes ''appropriate'' performance levels.

between the PLDs and the assessment even in a testing program that has a single form.

Writing PLDs following standard setting may promote their better alignment with the assessment. Where standards are set using the bookmark method, there may be some benefit to writing the PLDs following the standard setting (Hambleton, 2001). However, Delaware's standards were set using the bookmark method, and the PLDs were already in place (Delaware Department of Education, 2005). While we have not formally examined the alignment between Delaware's PLDs and its assessments, the Delaware process appears to have been comprehensive and robust. Ultimately, whether final PLDs are in place at the time of standard setting may matter less than whether there is a process in place to periodically review PLDs and ensure they are aligned with the competencies that are tested.

Ultimately, whether final PLDs are in place at the time of standard setting may matter less than whether there is a process in place to periodically review PLDs and ensure they are aligned with the competencies that are tested.

The primary advantage of having the PLDs developed in advance of the standard setting is that they will not be contaminated by the content of that particular administration of the test. On the other hand, we support writing or revising PLDs in conjunction with the standard setting. The standard setting process offers a rare opportunity for intense scrutiny of the PLDs in light of actual test items. Beginning the standard setting with general policy descriptions allows the panelists to internalize state policy on how good is good enough (Perie, 2004). However, having the actual test items at hand allows panelists the opportunity to reflect on expected performance with a specific, clear frame of reference and will likely result in PLDs that clearly describe what is needed for various levels of performance.

Sound Practices for Developing or Revising PLDs

Methods for developing or revising performance level descriptors have been outlined for specific assessments (Mills & Jaeger, 1998; Zwick, Senturk, Wang, & Loomis, 2001; Malagón, Rosenberg, & Winter, 2005). The procedures that have been outlined essentially follow these steps:

1. *Diverse panel*. Convene a diverse panel of subject matter experts and orient them to the task. Our experience suggests that having 4–8 panelists per assessment strikes a reasonable balance between ensuring a range of opinion and maintaining reasonable costs. In Delaware, about 70 professionals took part in rewriting the performance level descriptors in reading, writing, and mathematics across all grades (Delaware Department of Education, 2005).

2. *Item and task review*. Direct the panelists to review the items and tasks on the assessments.

3. *Relevant documents*. Provide the panelists with other relevant documents to review, such as the content standards, assessment frameworks/blueprints, and examples of student work, if appropriate.

4. *Generic descriptors*. Provide generic (cross-grade or cross-subject) descriptors for the panelists to review, if such descriptors exist. These generic descriptors may define student performance at pre-specified levels, such as basic, proficient, and advanced.

5. *Content and process analysis*. Conduct a content and process analysis of the items. The panelists should identify the academic knowledge needed to get the item correct (content) and what needs to be done with that knowledge (process) (Mills & Jaeger, 1998). In other words, *both the verbs and their objects matter.* For items with multiple score points, panelists need to identify these features for each score level (Malagón, Rosenberg, & Winter, 2005).

6. *Discussion of findings*. Conduct a discussion of panelists' findings, focused on relating their analysis of content and process to performance levels.

7. *Consensus*. Assist panelists in reaching consensus on descriptions of student performance at each level.

When coherence is an explicit goal of the PLD development or revision, some modification to the steps above may be in order. For example, the panelists may be grouped across grades so that assessments from adjacent grades are reviewed together. Reviewing the assessments from contiguous grades can highlight the differences in process demands and content emphasis between the grades. In both Michigan and Delaware, panelists were asked explicitly to identify what more is expected of students in one performance category and grade (e.g., met standards in grade 4) over students in the same performance category and the next lower grade (e.g., met standards in third grade) (Michigan Department of Education, 2006; Wise et al., 2005). When PLDs are developed or revised in conjunction with a standard setting using the bookmark method, using an ordered item booklet that spans the adjacent grades is a straightforward way to group items for content analysis.

Schulz, Lee, and Mullen (2005) used an entirely different analytic framework, focused on domains within a content area, to describe different levels of performance. In their study, which focused on the performance levels (Basic, Proficient, and Advanced) on the Grade 8 NAEP Mathematics test, teachers classified both secure and released items into content domains within mathematics. Next, curriculum experts identified the typical instructional timing for when the content reflected in an item was introduced and mastered. Both sets of ratings were reliable, and taken together, the ratings described domain-level patterns of mastery associated with the NAEP performance levels. The study begins to address the thorny question of how to determine overall performance when there is variation in performance by domain.

Format Examples

States have employed a range of formats for their performance level descriptors. In this section, we present a few examples. There are no

federal requirements for the format of PLDs. We believe that good practice in this area requires only that the PLDs be clear, but clarity can come in different formats. New Mexico, for example, employs short paragraphs of a few sentences in its high school reading PLDs (see Table 2). Pennsylvania's reading PLDs feature a single sentence with a series of bullets that embody that sentence (Table 3). For its PLDs, South Carolina has employed a two-column table, the cells of which are bulleted statements of what students at each level likely can and cannot do (Table 4). In Colorado and Nevada, the PLDs are organized by content standard within a subject (Table 5). As we pointed out when discussing the work of Schulz and his colleagues (2005), this organization highlights information at a finer level but raises the challenge of how to set overall cut scores when there is a mix of levels.[14] Additionally, as seen in the Colorado example, the statement, ''No evidence of this standard at this performance level,'' appears often in the lower proficiency level descriptions, which can be problematic from a descriptive standpoint.

Good practice in developing PLDs requires that the PLDs be clear, but clarity can come in different formats.

Table 2. New Mexico High School Reading Performance Level Descriptors

Advanced
High school students performing at the Advanced Level in reading demonstrate a highly developed and comprehensive understanding of various genres. They use a wide range of sophisticated strategies to make critical evaluations and understand advanced literary devices. These students read challenging text and are able to critically interpret and evaluate literature, language, and ideas. They apply significant understanding to develop hypotheses and perform critical analyses of the complex connections among texts. Advanced Level students recognize subtle inferences and differentiate among various levels of reading.

Proficient
High school students performing at the Proficient Level demonstrate a developed understanding of various genres. These students are able to draw and support conclusions using textual evidence. They identify, respond to, and evaluate problems and solutions. These students are able to recognize and evaluate a writer's position within a text. They also differentiate among literal, connotative, and figurative meanings and are able to make logical inferences. These students analyze information and interpret critical details. Proficient Level students communicate and organize their ideas coherently, demonstrating what is relevant and accurate.

Nearing Proficient
High school students performing at the Nearing Proficient Level demonstrate a developing understanding of various genres. They are able to make logical, though limited, connections. These students have the ability to recognize interpretations; they also understand the significance of problems and solutions presented. Nearing Proficiency Level students respond to the text at a literal level, exhibit some skill in making inferences, yet make some errors when recalling facts.

Source: New Mexico Department of Education (2003)

[14] For more information, see Human Resources Research Organization (2006). At the time of writing, this resource is a growing online database of performance level descriptors from each of the states at grades 4 and 8 and at high school in a standardized format.

Table 3. Pennsylvania Grade 3 Reading Performance Level Descriptors

Below Basic A student scoring at the below basic level demonstrates competency with below grade-level text only and requires extensive support to comprehend and interpret fiction and nonfiction.

Basic A student scoring at the basic level generally utilizes some reading strategies to comprehend grade-level appropriate fiction and nonfiction:

- Identifies some word meanings, including synonyms and antonyms for common words, using context clues
- Identifies details in support of a conclusion
- Identifies stated main ideas
- Attempts to summarize text
- Attempts to make within or among text-to-text connections
- Identifies purpose of text (e.g., narrative)
- Identifies some literary elements (e.g., character)
- Locates headings and subheadings in text
- Recognizes simple organizational patterns of text (e.g., sequencing and comparison/contrast)
- Recognizes that authors use language in different ways to communicate meaning
- Identifies factual statements
- Recognizes graphics and charts

Proficient A student scoring at the proficient level routinely utilizes a variety of reading strategies to comprehend and interpret grade-level appropriate fiction and nonfiction:

- Identifies word meanings, including synonyms and antonyms, using context clues and word parts
- Makes inferences and draws conclusions, using textual support
- Identifies stated or implied main ideas and relevant details
- Summarizes text
- Makes within and among text-to-text connections
- Identifies purpose of text (e.g., narrative and informational)
- Identifies literary elements (e.g., character, setting and plot)
- Identifies figurative language (e.g., personification)
- Identifies fact and opinion and the use of exaggeration (bias) in nonfiction
- Identifies organizational patterns of text (e.g., sequencing and comparison/contrast)
- Interprets graphics, charts, and headings

Advanced A student scoring at the advanced level consistently utilizes sophisticated strategies to comprehend and interpret complex fiction and nonfiction:

- Identifies word meanings and shades of meaning, using context as support
- Makes inferences and draws conclusions, using textual support
- Relates supporting details to main idea
- Effectively summarizes all ideas within text
- Describes within and among text-to-text connections
- Explains purpose of text (e.g., narrative)
- Explains the use of figurative language (e.g., personification and simile) and literary elements (e.g., character)
- Explains the use of fact and opinion and exaggeration (bias) in nonfiction
- Identifies and explains organizational patterns of text (e.g., sequencing and comparison/contrast)
- Applies information in graphics, charts, and headings to support text

Source: Pennsylvania Department of Education (n. d.)

Table 4. South Carolina Grade 5 English Language Arts Performance Level Descriptors

Below Basic	
What *below basic* students likely can do: • paraphrase main ideas in a variety of texts • identify explicit details • read informational and literary texts • make simple inferences when abundant clues are present • use word-matching strategies to answer questions	**What *below basic* students likely cannot do:** • focus on pieces of longer and denser text • discriminate among a limited number of details to select to the most relevant detail • process multiple details • make inferences when details are scattered or heavily embedded in text • use relevant details when answering constructed-response items
Basic	
What *basic* students likely can do that *below basic* students likely cannot do: • identify and analyze multiple details • make inferences when pictures support the text • read longer informational and literary texts • identify appropriate research sources • recognize and categorize synonyms • identify character traits	**What *basic* students likely cannot do:** • analyze and interpret figurative language • provide an interpretation that goes beyond the text • use multiple reading strategies simultaneously • use text structures to locate relevant information • analyze character traits
Proficient	
What *proficient* students likely can do that *basic* students likely cannot do: • use multiple reading strategies simultaneously • use context to determine the meaning of words • move through scattered details to comprehend longer and more complex texts (synthesize text) • interpret figurative language • analyze character traits	**What *proficient* students likely cannot do:** • use a dictionary and context clues to determine the meaning of multiple-meaning words in increasingly complex texts
Advanced	
What *advanced* students likely can do that *proficient* students likely cannot do: • locate relevant details in increasingly complex texts • use a dictionary and context clues to analyze the multiple meanings of a word in increasingly more complex texts	

Source: South Carolina Department of Education (2005)

Table 5. Colorado Grade 9 Mathematics Performance Level Descriptors

Advanced

Standard 1
Students demonstrate exceptional use of number sense and use of numbers by
- Recognizing the properties of exponents

Students may also demonstrate exceptional use of number sense and use of numbers by
- Using properties of exponents to express ratios between two numbers written in scientific notation
- Using the properties of exponents to apply the operation ''to the power of''

Standard 2
Students demonstrate exceptional use of algebraic methods to explore, model, and describe patterns and functions by
- Representing functional relationships in multiple ways
- Expressing the perimeter of geometric figures algebraically
- Determining the solution to simple systems of equations using graphing
- Solving problems using algebraic methods
- Modeling real-world situations using equations

Students may also demonstrate exceptional use of algebraic methods to explore, model, and describe patterns and functions by
- Modeling real-world situations using patterns and equations
- Solving simple systems of equations using algebraic methods
- Identifying and interpreting x- and y-intercepts in the context of problems
- Solving problems involving comparison of rates
- Solving for the independent variable when given the dependent variable

Standard 3
Students demonstrate exceptional use of data collection and analysis, statistics, and probability by
- Determining measures of central tendency from graphed data
- Determining the effects of additional data on measures of variability and central tendency
- Drawing lines of best fit to make predictions about data

Students may also demonstrate exceptional use of data collection and analysis, statistics, and probability by
- Describing how data can be used to support more than one position
- Determining quartiles
- Determining the probability of dependent and independent events
- Determining appropriate measures of central tendency from given data in the context of problems
- Using permutations to solve real-world problems
- Applying understanding of the relationship among measures of central tendency
- Determining equations to represent lines of best fit
- Interpreting, interpolating, and extrapolating using lines of best fit in real-world situations
- Interpreting measures of variability in problem-solving situations
- Interpreting slope in the context of problems

Table 5. Colorado Grade 9 Mathematics Performance Level Descriptors *(cont.)*

Advanced (cont.)

Standard 4
Students demonstrate exceptional use of geometric concepts, properties, and relationships by
- Demonstrating how changing dimensions and shapes of simple figures affects their perimeters
- Calculating the volume of simple geometric solids
- Applying the concept of slope to locate points on a coordinate grid
- Recognizing the relationship between the areas and sides of simple figures
- Determining how a change in the dimensions or shape of a figure affects perimeter
- Applying the Pythagorean theorem in real-world situations
- Recognizing angle relationships within figures

Students may also demonstrate exceptional use of geometric concepts, properties, and relationships by
- Determining maximum and minimum perimeter values when the dimensions of figures are changed
- Representing irrational numbers and their squares geometrically
- Explaining the relationship between the areas and sides of simple figures

Standard 5
Students demonstrate exceptional use of a variety of tools and techniques to measure by
- Modeling rate of change in real-world situations involving different units
- Using appropriate measurement tools and scale factors to calculate rates of change in multistep problems
- Explaining methods for finding the area of triangles using the Pythagorean theorem
- Describing the change in volume of a shape that results from changing one attribute of that shape

Students may also demonstrate exceptional use of a variety of tools and techniques to measure by
- Calculating and justifying solutions to geometric problems requiring the use of the Pythagorean theorem
- Using measurements to indirectly solve problems involving surface area

Standard 6
Students demonstrate exceptional use of computational techniques in problem-solving situations by
- Converting from one set of units to another
- Selecting and using operations in problem-solving situations involving whole numbers and percents

Students may also demonstrate exceptional use of computational techniques in problem-solving situations by
- Selecting and using operations in multistep problems involving percents and proportional thinking

Table 5. Colorado Grade 9 Mathematics Performance Level Descriptors *(cont.)*

Proficient

Standard 1
Students demonstrate use of number sense and use of numbers by
- Estimating the reasonableness of solutions involving rational numbers
- Translating numbers from standard notation to scientific notation

Standard 2
Students demonstrate use of algebraic methods to explore, model, and describe patterns and functions by
- Converting from one functional representation to another
- Representing functional relationships using an equation or table
- Evaluating formulas
- Interpreting graphical representations of real-world situations
- Graphing functional relationships

Standard 3
Students demonstrate use of data collection and analysis, statistics, and probability by
- Using appropriate data displays to represent and describe sets of data
- Determining the probability of identified events using the sample spaces
- Describing how data can be used to support claims

Standard 4
Students demonstrate use of geometric concepts, properties, and relationships by
- Using coordinate geometry to solve problems involving the midpoint of a segment
- Using transformation concepts to identify relationships between parts of figures
- Applying knowledge of perimeters in problem-solving situations

Standard 5
Students demonstrate use of a variety of tools and techniques to measure by
- Using appropriate measurement tools and scale factors to find unknown measurements

Standard 6
Students demonstrate use of computational techniques in problem-solving situations by
- Using proportional thinking in problem-solving situations
- Computing using rational numbers
- Selecting and using operations to solve problems involving rational numbers and percents

Partially Proficient

Standard 1
No evidence of this standard at this performance level.

Standard 2
Students demonstrate limited use of algebraic methods to explore, model, and describe patterns and functions by
- Translating written relationships into equations
- Using graphs to identify the maximum and minimum within given domains

Table 5. Colorado Grade 9 Mathematics Performance Level Descriptors *(cont.)*

Partially Proficient (cont.)

Standard 3
Students demonstrate limited use of data collection and analysis, statistics, and probability by
- Using counting strategies to determine the possible outcomes of a process

Standard 4
No evidence of this standard at this performance level.

Standard 5
No evidence of this standard at this performance level.

Standard 6
Students demonstrate limited use of computational techniques in problem-solving situations by
- Computing with integers

Unsatisfactory

Standard 1
No evidence of this standard at this performance level.

Standard 2
Students demonstrate minimal use of algebraic methods to explore, model, and describe patterns and functions by
- Working backwards to solve problems

Standard 3
Students demonstrate minimal use of data collection and analysis, statistics, and probability by
- Reading, interpreting, and comparing displays of data

Standard 4
No evidence of this standard at this performance level.

Standard 5
No evidence of this standard at this performance level.

Standard 6
Students demonstrate minimal use of computational techniques in problem-solving situations by
- Computing with whole numbers in basic single-step problems

Source: Colorado Department of Education (2002)

Summary

Setting coherent standards across grades requires the blending of traditional standard setting practices with new considerations that reflect policy choices, content progressions, and empirical results. The key decision points for a state embarking on setting coherent standards are summarized in Appendix A. These considerations are specific to setting coherent standards across grades, summarizing much of the content of this paper. They draw from state experiences in setting standards across grades. Selected state examples are in Appendix B.

Both Schafer (2005) and Hambleton (2001) have proposed general criteria for standard setting, Hambleton from a grade-by-grade perspective and Schafer from the perspective of cross-grade consistency with a focus on the needs of state departments of education. Both papers are excellent resources for planning standard setting activities. Though still evolving, state experiences with setting coherent standards point to four key decision points for planning cross-grade standard setting:

1. *Panel configuration.* The panelists should include educators with appropriate teaching experience. If panels set cut scores for a single grade, then the panels should include educators with teaching experience at adjacent grades. If panels set cut scores for multiple grades, then educators with teaching experience across the relevant grades should be included.

2. *Basis for cut scores.* State planners will need to choose a course that reflects the desired balance between panelist judgment and empirical results. Coherence does not require a particular method; however, at the end of the process, the cut scores need to relate logically to one another, in terms of both content coverage and impact.

3. *Articulation.* Smoothing the cut scores can reinforce a logical (quantitative) progression across grades. The performance standards need to have just as logical a progression through the content, so well-articulated performance level descriptors are essential.

4. *PLD development.* Beginning with generic PLDs and developing more specific ones at the standard setting ensures that the PLDs reflect the test content. On the other hand, specific PLDs promote a common set of panelist expectations. Provisions should be made for external or state department-based content and assessment experts to refine the PLDs after the meeting based on panelist comments.

References

Buckendahl, C., Huynh, H., Siskind, T., & Saunders, J. (2005). A case of vertically moderated standard setting for a state science assessment program. *Applied Measurement in Education, 18*(1), 83–98.

Cizek, G. J. (2001). *Setting performance standards: Concepts, methods, and perspectives.* Mahwah, NJ: Lawrence Erlbaum.

Cizek, G. J. (2005). Adapting testing technology to serve accountability aims: The case of vertically moderated standard setting. *Applied Measurement in Education, 18*(1), 1–9.

Colorado Department of Education. (2002). *Colorado Student Assessment Program: Mathematics grade 9 performance level descriptors*. Retrieved January 25, 2006, from http://www.cde.state.co.us/cdeassess/csap/PLD/as_g9MathPLD.htm

CTB/McGraw-Hill. (2005). *Missouri Assessment Program: Sections A, B, F, G, and I of the bookmark standard setting technical report for grades 3, 4, 5, 6, 7, 8, and 11 communications arts & grades 3, 4, 5, 6, 7, 8, and 10 mathematics*. Draft submitted to the Missouri Department of Education.

Data Recognition Corporation. (2005a). *Alaska Comprehensive System of Student Assessment technical report*. Juneau, AK: Alaska Department of Education.

Data Recognition Corporation. (2005b). *Performance levels validation report*. Harrisburg, PA: Pennsylvania Department of Education.

Delaware Department of Education. (2005). *A summary report and recommendations to the Delaware State Board of Education for revisiting, reviewing, and establishing performance standards for the Delaware Student Testing Program reading, writing, and mathematics*. Dover, DE: Assessment and Analysis Work Group. Retrieved January 24, 2006, from http://www.doe.k12.de.us/AAB/Merged%20report%20of%20cut%20score%20review%20Oct%202005.pdf

Hambleton, R. K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. J Cizek (Ed.), *Setting performance standards: Concepts, methods and perspectives* (pp. 89–116). Mahwah, NJ: Lawrence Erlbaum.

Hansche, L. N. (1998). *Meeting the requirements of Title I: Handbook for the development of performance standards.* Washington, DC: U.S. Department of Education.

Human Resources Research Organization. (2006). *Knowing what students need to know: Performance level descriptions (PLD).* Alexandria, VA: Author. Retrieved February 14, 2006, from http://new.humrro.org/PLD/

Huynh, H., Barton, K. E., Meyer, J. P., Porchea, S., & Gallant, D. (2005). Consistency and predictive nature of vertically moderated standards for South Carolina's 1999 Palmetto Achievement Challenge Tests of language arts and mathematics. *Applied Measurement in Education, 18*(1), 115–128.

Huynh, H., Meyer, J. P., & Barton, K. (2000). *Technical documentation for the 1999 Palmetto Achievement Challenge Tests of English language arts and mathematics, grades three through eight*. Columbia, SC: South Carolina Department of Education, Office of Assessment. Retrieved March 22, 2006, from http://www.myscschools.com/offices/assessment//Publications/1999_Pact_document.doc.

Huynh, H., & Schneider, C. (2005). Vertically moderated standards: Background, assumptions, and practices. *Applied Measurement in Education, 18*(1), 99–113.

Lewis, D. M. (2002, April). *Standard setting with vertical scales*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Lewis, D. M., & Haug, C. A. (2005). Aligning policy and methodology to achieve consistent across-grade performance standards. *Applied Measurement in Education, 18*(1), 11–34.

Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996, June). Standard setting: A bookmark approach. Paper presented at the Council of Chief State School Officers National Conference on Large-Scale Assessment, Phoenix, AZ.

Linn, R. L. (2003). Performance standards: Utility for different uses of assessments. *Education Policy Analysis Archives, 11*(31). Retrieved February 11, 2006, from http://epaa.asu.edu/epaa/v11n31/

Lissitz, R. W., & Huynh, H. (2003). Vertical equating for state assessments: issues and solutions in determination of adequate yearly progress and school accountability. *Practical Assessment, Research & Evaluation, 8*(10). Retrieved February 6, 2006, from http://PAREonline.net/getvn.asp?v = 8&n = 10

Malagón, M. H., Rosenberg, M. B., & Winter, P. C. (2005). *Developing aligned performance level descriptors for the English language development assessment K-2 inventories*. Washington, DC: Council of Chief State School Officers.

Michigan Department of Education. (2006). *2006 MEAP standard setting: Mathematics, reading, writing, science, and social studies*. Lansing, MI: Assessment and Evaluation Services.

Mills, C. N., & Jaeger, R. M. (1998). Creating descriptions of desired student achievement when setting performance standards. In L. N. Hansche, *Meeting the requirements of Title I: Handbook for the development of performance standards.* Washington, DC: U.S. Department of Education.

Mitzel, H. C. (2005). *Consistency for state achievement standards under NCLB*. Washington, DC: Council of Chief State School Officers.

New Mexico Department of Education. (2003). *Reading performance level descriptors*. Retrieved January 25, 2006, from http://www.ped.state.nm.us/div/acc.assess/assess/dl/final.reading.level.descriptors.1210031.pdf

Ohio Department of Education. (2005). *Ohio Achievement Tests: Spring 2005 standard setting (Setting standards in grades 4–8 reading, grades 3–8 mathematics, and grade 4 writing: Technical report).* Columbus, OH: Author.

Pennsylvania Department of Education. (n.d.) *Grade 3 reading performance level descriptors*. Harrisburg, PA: Author. Retrieved January 25, 2006, from http://www.pde.state.pa.us/a_and_t/lib/a_and_t/Grade3ReadingPerformanceLevelDescriptors.pdf

Perie, M. (2004, July). *Considerations for standard setting on K–12 assessments*. Unpublished presentation. Educational Testing Service.

Ravitch, D. (2006, January 5). National standards: ''50 Standards for 50 States'' is a formula for incoherence and obfuscation. *Education Week*. Retrieved March 9, 2006, from http://www.edweek.org/ew/articles/2006/01/05/17ravitch.h25.html

Schafer, W. D. (2005). Criteria for standard setting from the sponsor's perspective. *Applied Measurement in Education, 18*(1), 61–81.

Schulz, E. M., Lee, W. C., & Mullen, K. (2005). A domain-level approach to describing growth in achievement. *Journal of Educational Measurement, 42*(1), 1–26.

South Carolina Department of Education. (2005). *The development of performance level descriptors for the English language arts and mathematics PACT tests*. Columbia, SC: Author.

U.S. Department of Education. (2004, April 28). *Standards and assessments peer review guidance: Information and examples for meeting requirements of the No Child Left Behind Act of 2001*. Washington, DC: Office of Elementary and Secondary Education. Retrieved November 4, 2005, from http://www.ed.gov/policy/elsec/guid/saaprguidance.doc

U.S. Department of Education, National Center for Education Statistics. (2005). *The condition of education 2005* (NCES 2005–094). Washington, DC: U.S. Government Printing Office.

Williams, A., Blank, R., Cavell, L., & Toye, C. (2005). *State education indicators with a focus on Title I: 2001–02*. Washington, DC: U.S. Department of Education.

Wise, L. L., Zhang, L., Winter, P., Taylor, L., & Becker, D. E. (2005). *Vertical alignment of grade-level expectations for student achievement: Report of a pilot study.* Alexandria, VA: Human Resources Research Organization (HumRRO).

Zhang, L. (2005, October). *The Delaware experience: Revisit, review, and establish performance standards (cut scores) for reading, writing, and mathematics*. Presentation at the State Collaborative on Assessment and Student Standards (SCASS) Technical Issues in Large-Scale Assessment consortium meeting, Charleston, SC.

Zwick, R., Senturk, D., Wang, J., & Loomis, S. C. (2001). An investigation of alternative methods for item mapping in the National Assessment of Educational Progress. *Educational Measurement: Issues and Practice, 20*(2), 15–25.

Appendix A: Key Decision Points for Cross-Grade Standard Setting

Panel configuration

1. *Panels set cut scores for a single grade.* If there will be separate panels for each grade level of the test, each panel should include educators with experience teaching adjacent grades. Provisions should be made for sharing information across panels regarding impact data of proposed cut scores.
2. *Panels set cut scores for multiple grades.* Decisions need to be made about the order of grade levels in the process and how to incorporate information from each grade's proposed standards in the process. While this panel configuration is the most logical for cross-grade standard setting, it requires careful thought and preparation for each step in the process. Ideally, panelists will have familiarity and experience with each grade for which they will be setting standards.

Approach

1. *Primarily judgment based.* States that are developing a new testing program across grades will not have existing cut scores or impact data for setting preliminary cut scores on their tests. If no starting information will be used to guide panelists, careful attention needs to be paid to sharing potential impact data, such as field test data, across grades and creating a common understanding of expected performance.
2. *Primarily data-based.* Statistical interpolation between the lowest and highest grade tested, with no adjustment based on test review, falls into this category. While this technique ensures a coherent set of cut scores, it may not result in standards that are coherent in terms of content coverage and grade-level expectations.
3. *Combined data and judgment.* Most states have existing cut scores for some of the tested grade levels. Impact data can be used to recommend preliminary standards to panelists based on the existing cut scores. In this case, panelists have to clearly understand the degree to which they can modify the preliminary cut scores and should be encouraged to tie their modifications to specific content considerations. Another option is to provide reference points for panelists, showing where cut scores would be located if the impact were the same as the impact in a reference year, usually the previous year.

Articulation across grades

1. *Smoothing.* Smoothing of cut scores across grades can be conducted at various points in the standard setting process. Providing panelists with preliminary cut scores at the beginning of the process, showing where proposed cut scores fall on a vertical scale during the process, and having a subset of panelists across grades adjust cut scores after the panels have completed their tasks are techniques that have been used alone and in combination.
2. *Presenting cross-grade data.* If preliminary cut scores are used, impact data across grades can be discussed by the entire group at the start of the session to foster a shared understanding of student performance. As the standard setting meeting progresses, panelists can review the impact of proposed cut scores from other grades to help ground their judgments.
3. *Presenting cross-grade content and test information.* Panelists will have a better understanding of how the content progresses across grades if they review materials from other grade levels. These materials may include content standards and test blueprints for adjacent (or all) grade levels, and PLDs for all grade levels, if they have been drafted prior to standard setting.
4. *Cross-grade discussions.* At various points in the standard setting process, it may be useful for panels setting cut scores in adjacent grades (or all grades) to convene in order to discuss the rationales for their judgments and to ask each other questions about content across the grades. If this technique is used, it must be carefully structured (and, in some cases, facilitated) to maintain focused, on-task discussion.

PLD development

1. *Start with generic PLDs.* Some states have started their standard setting sessions with policy descriptors that are common across content areas and grade levels or with content-specific descriptors that are common across grade levels. In this case, content- and grade-level descriptors may be drafted by the standard setting panelists, a subset of the panelists, or a separate group of content experts. The standard-setting procedure must elicit information that will assist in writing the descriptors.
2. *Start with specific PLDs.* If grade-level/content-specific PLDs are developed before the standard-setting meeting, panelists have a common set of expectations to guide their judgments. Provisions should be made to refine the PLDs after the meeting based on panelists' comments.

Appendix B: Selected State Examples of Cross-Grade Standard Setting Procedures (Mathematics and Language Arts)

State	Panel Configuration[1]	Approach	Articulation across Grades	PLD Development
Alaska	Multiple grades	Combination – preliminary cut scores presented to panel	• Preliminary cut scores were set to conform to grade 6 percentages in each level. • Standard setting included cross-grade discussion of impact data.	Grade/content specific PLDs were used in the standard-setting meeting. Suggestions for refining PLDs were made by panelists, and PLDs were refined as needed after the sessions.
Delaware	Single grade	Combination – preliminary cut scores presented to panel	For grades 3, 5, 8, and 10 • Panelists were shown the location and impact data for existing cut scores and discussed patterns of performance across grades. • Panelists were shown the position of the cut scores on the vertical scale. • Recommended cut scores were adjusted as needed based on patterns of performance and location on the vertical scale. For grades 2, 4, 6, 7, and 9 • Preliminary cut scores were set using interpolation and extrapolation from cut scores set earlier for grades 3, 5, 8, and 10; they were smoothed if needed, based on patterns of performance and location on the vertical scale. Panelists discussed patterns of performance across grades. • Panelists were shown the position of the cut scores on the vertical scale. • Recommended cut scores were adjusted as needed based on patterns of performance and location on the vertical scale.	Grade/content specific PLDs were used in the standard-setting meeting. PLDs were refined as needed after the sessions.

[1] **Panel Configuration**: whether each panel set cut scores on multiple grades or a single grade.
Approach: whether and how test data were used in setting standards.
Articulation: techniques used (a) to prepare for, (b) during, or (c) after a standard setting session to promote cross-grade coherence in the panel results.
PLD Development: type and timing of PLDs developed/used for standard setting.

State	Panel Configuration[1]	Approach	Articulation across Grades	PLD Development
Michigan	Multiple grades	Combination – preliminary cut scores presented to panel	• For base grades, reference standards were based on prior year's performance to assist panelists in understanding the impact of their recommendations in terms of changes in impact on prior years' results. • For intermediate grades, reference cut scores were interpolated/extrapolated from base grades. • Standard setting included full group discussion of impact data.	Grade/content specific PLDs were used in the standard-setting meeting. Suggestions for refining PLDs were made by panelists, and PLDs were refined as needed after the sessions.
Missouri	Single grade	Combination – boundaries set on permissible cut scores	• Boundaries were set for all grades based on NAEP and prior state assessment performance. • Standard setting included cross-grade discussion of impact data.	Content specific, not grade-level specific, PLDs were used in the standard-setting meeting. Grade/content specific PLDs were developed based on the results of the sessions.
Ohio	Multiple grades	Primarily judgmental	• Cut scores were set on "anchor grades" first; the impact data from anchor cut scores were used to set preliminary cut scores for intermediate grades. • Standard setting included cross-grade discussion of impact data.	Grade/content specific PLDs were used in the standard-setting meeting.
Pennsylvania	Multiple grades	Combination – preliminary cut scores presented to panel	• Preliminary cut scores were based on exponential growth functions across grades (extrapolated for grade 3). • Standard setting included full-group discussion of impact data. • Recommendations falling outside standard error bands were adjusted to conform to the limits of the bands, with recommendations above the band adjusted to the high level of the band and recommendations below the band adjusted to the low level of the band.	Grade/content specific PLDs were used in the standard-setting meeting.

State	Panel Configuration[1]	Approach	Articulation across Grades	PLD Development
South Carolina	Grades 3 and 8 • Single grade panels Grades 2–7 • N/A	Grades 3 and 8 • Primarily judgmental Grades 2, 4, 5, 6, and 7 • Primarily data-based – for grades 4, 5, 6, and 7, statistical interpolation between cut scores set for grades 3 and 8; for grade 2, extrapolation from grade 3 cut score	N/A	Generic PLDs were used in the standard-setting meeting. Grade/content specific PLDs were developed after cut scores were set.
West Virginia	Single grade	Primarily judgmental	Standard setting included cross-grade discussion of impact data.	Grade/content specific PLDs were used in the standard-setting meeting.

[1] **Panel Configuration**: whether each panel set cut scores on multiple grades or a single grade.
Approach: whether and how test data were used in setting standards.
Articulation: techniques used (a) to prepare for, (b) during, or (c) after a standard setting session to promote cross-grade coherence in the panel results.
PLD Development: type and timing of PLDs developed/used for standard setting.

Appendix C: South Carolina and Colorado Test Results from Standard-Setting Year to 2005

South Carolina Math

Percentages at Each Performance Level

1999

Grade	Below Basic	Basic	Proficient	Advanced
3	44	38	13	5
4	45	37	13	5
5	47	37	12	4
6	47	37	11	5
7	48	36	11	5
8	49	36	10	5

2005

Grade	Below Basic	Basic	Proficient	Advanced
3	17	53	21	10
4	21	38	26	14
5	23	45	18	15
6	21	40	25	14
7	29	39	18	15
8	34	43	15	8

South Carolina English Language Arts

Percentages at Each Performance Level

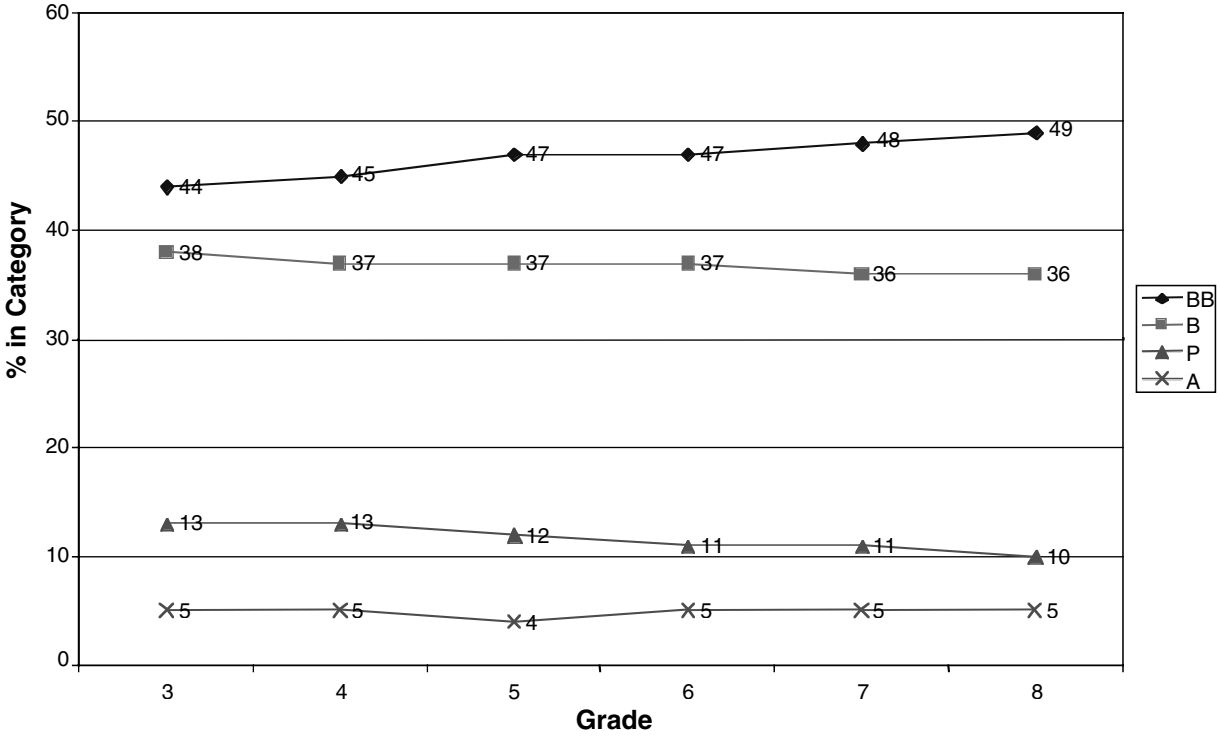
1999

Grade	Below Basic	Basic	Proficient	Advanced
3	35	37	26	2
4	35	37	26	2
5	35	39	24	2
6	37	39	21	3
7	37	39	21	3
8	38	41	19	3

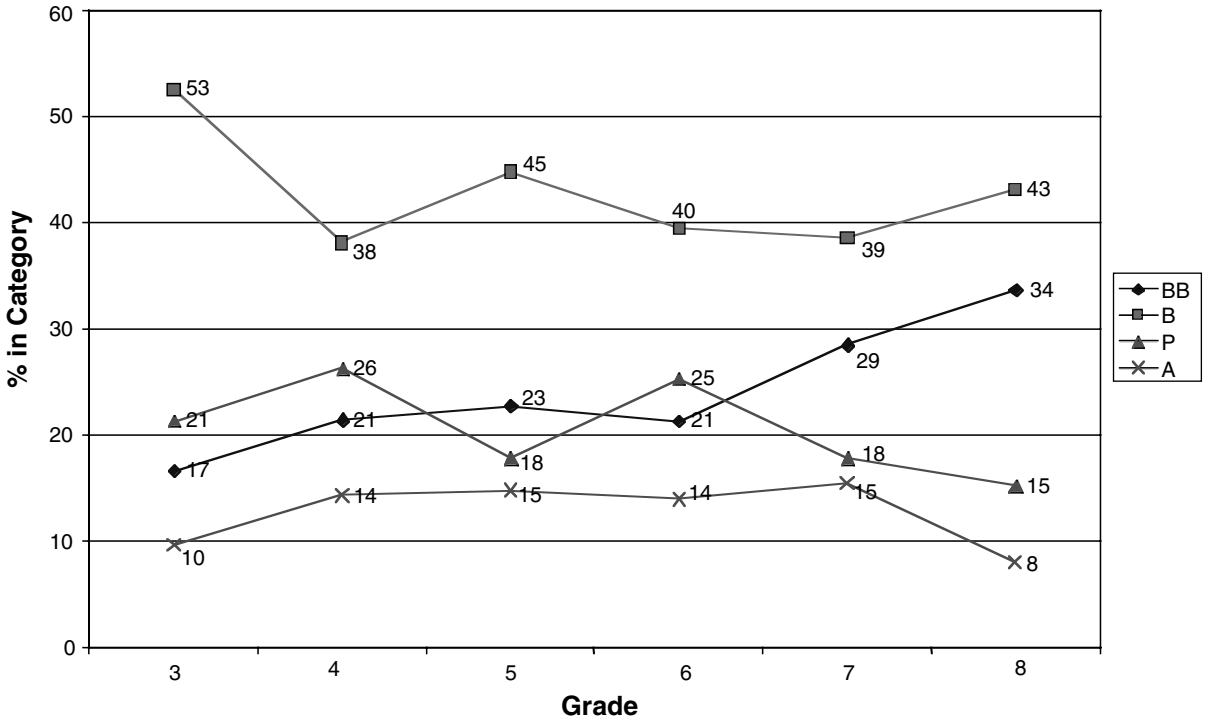
2005

Grade	Below Basic	Basic	Proficient	Advanced
3	13	30	48	9
4	20	43	34	3
5	23	47	28	2
6	37	36	22	5
7	29	47	22	3
8	25	45	24	6

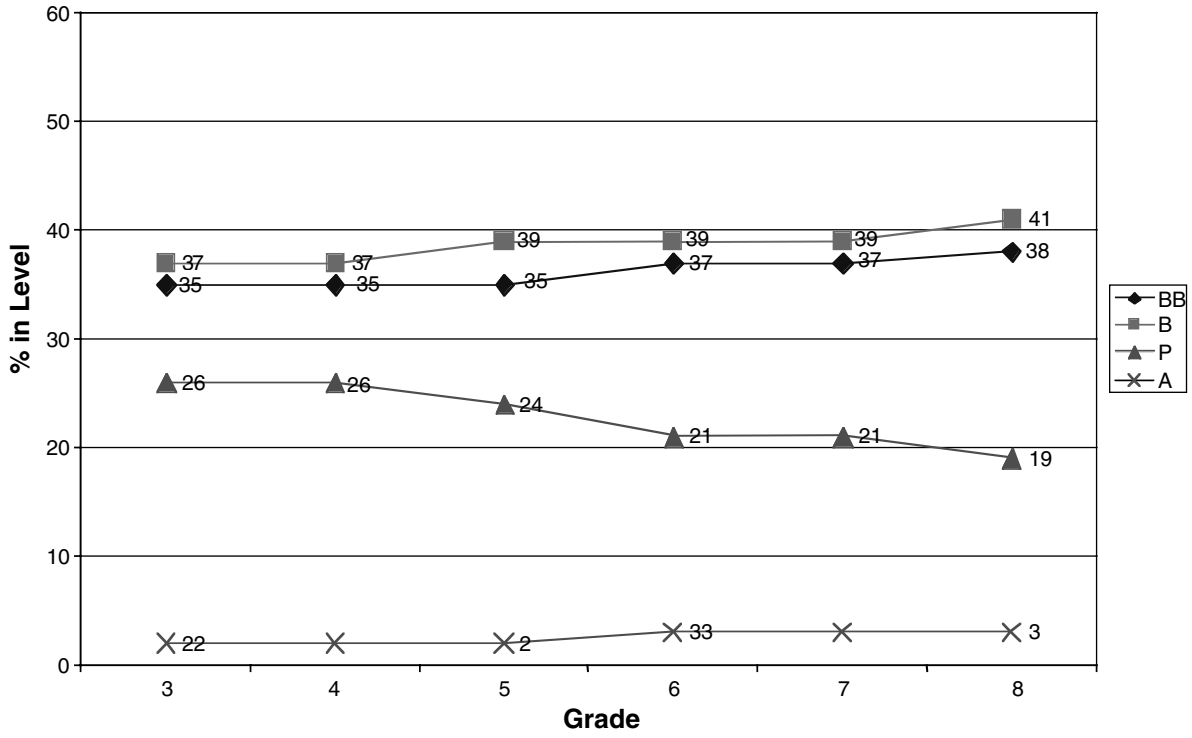
SC Math 1999



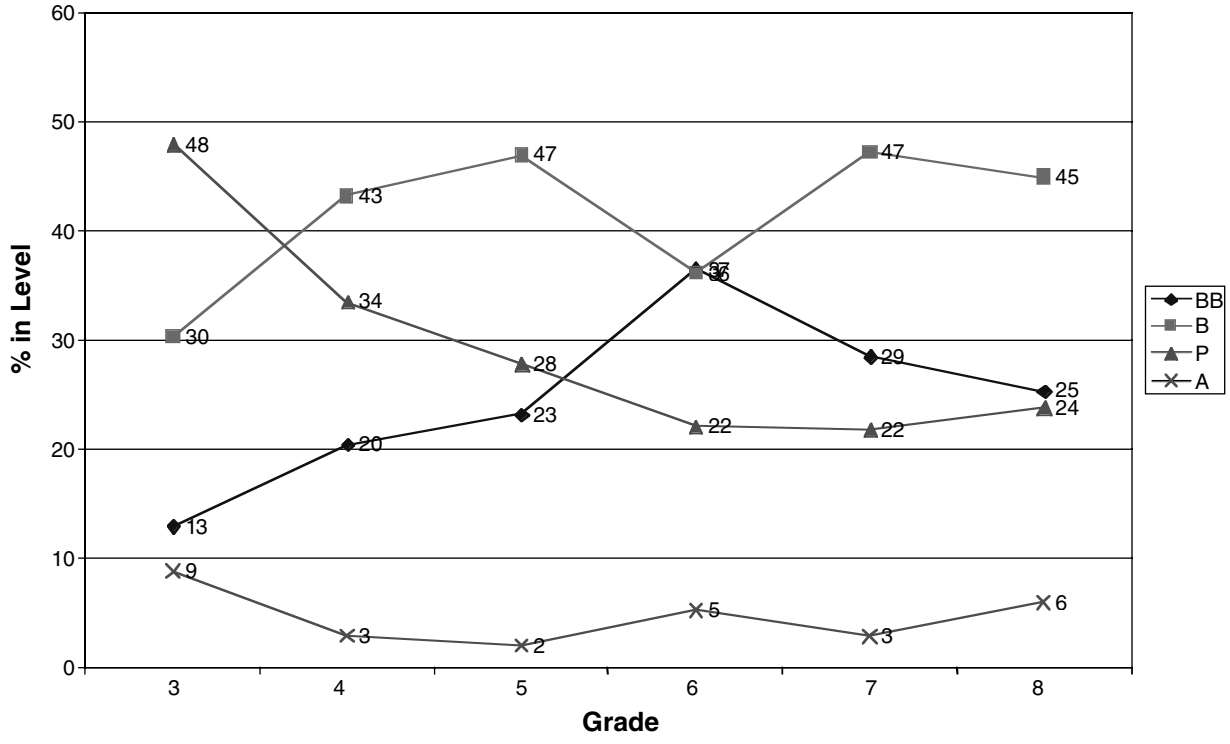
SC Math 2005



SC English Language Arts 1999



SC English Language Arts 2005



Colorado Math

Percentages at Each Performance Level

2002

Grade	Unsatisfactory	Partially Proficient	Proficient	Advanced	No Score Reported
05	12	31	35	20	2
06	16	30	35	16	3
07	21	36	27	11	4
08	26	31	26	13	4
09	34	29	22	9	5
10	31	37	24	3	5

2005

Grade	Unsatisfactory	Partially Proficient	Proficient	Advanced	No Score Reported
05	10	26	36	27	1
06	14	29	34	22	1
07	16	36	28	18	2
08	23	31	29	15	2
09	33	30	23	10	4
10	32	35	25	5	3

Colorado Writing

Percentages at Each Performance Level

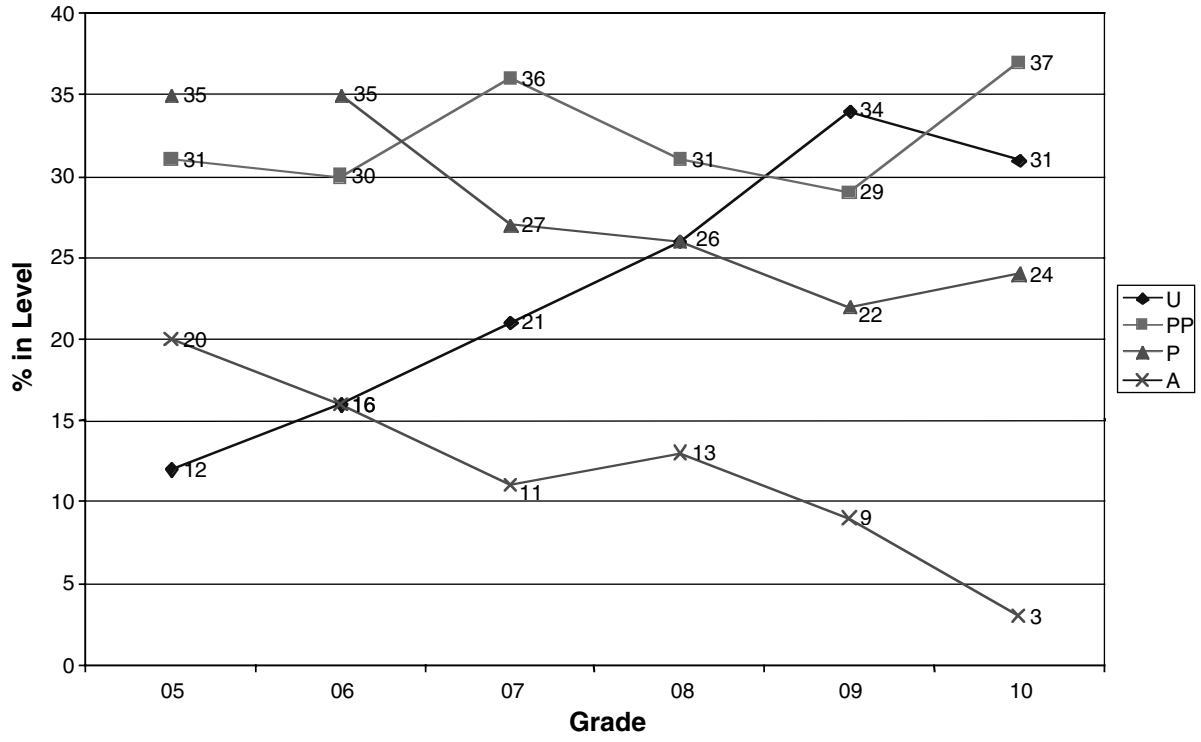
2002

Grade	Unsatisfactory	Partially Proficient	Proficient	Advanced	No Score Reported
03	7	40	43	8	2
04	8	40	42	8	1
05	7	39	42	8	3
06	7	39	42	8	3
07	4	42	42	8	4
08	5	41	42	8	4
09	6	40	41	8	5
10	6	39	42	8	5

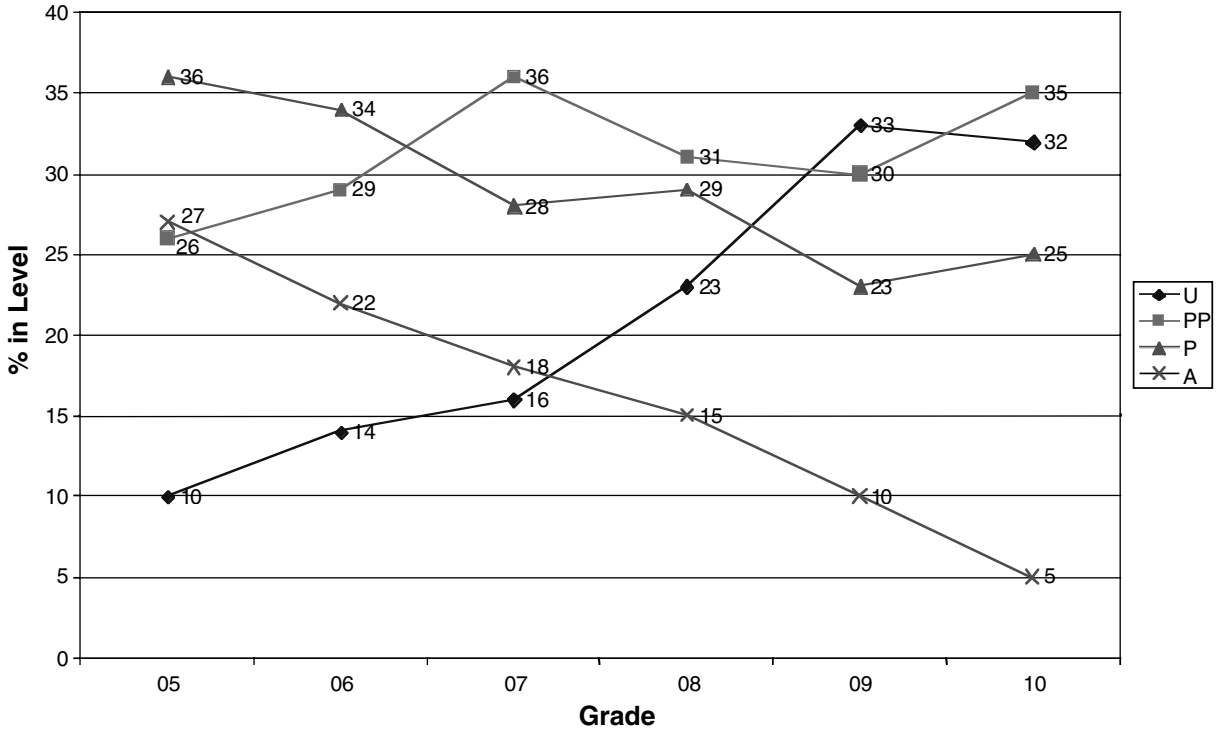
2005

Grade	Unsatisfactory	Partially Proficient	Proficient	Advanced	No Score Reported
03	5	38	47	9	1
04	8	39	43	9	1
05	5	37	48	10	1
06	5	34	48	11	2
07	5	37	44	12	2
08	5	42	43	9	2
09	5	40	44	8	4
10	7	40	43	7	4

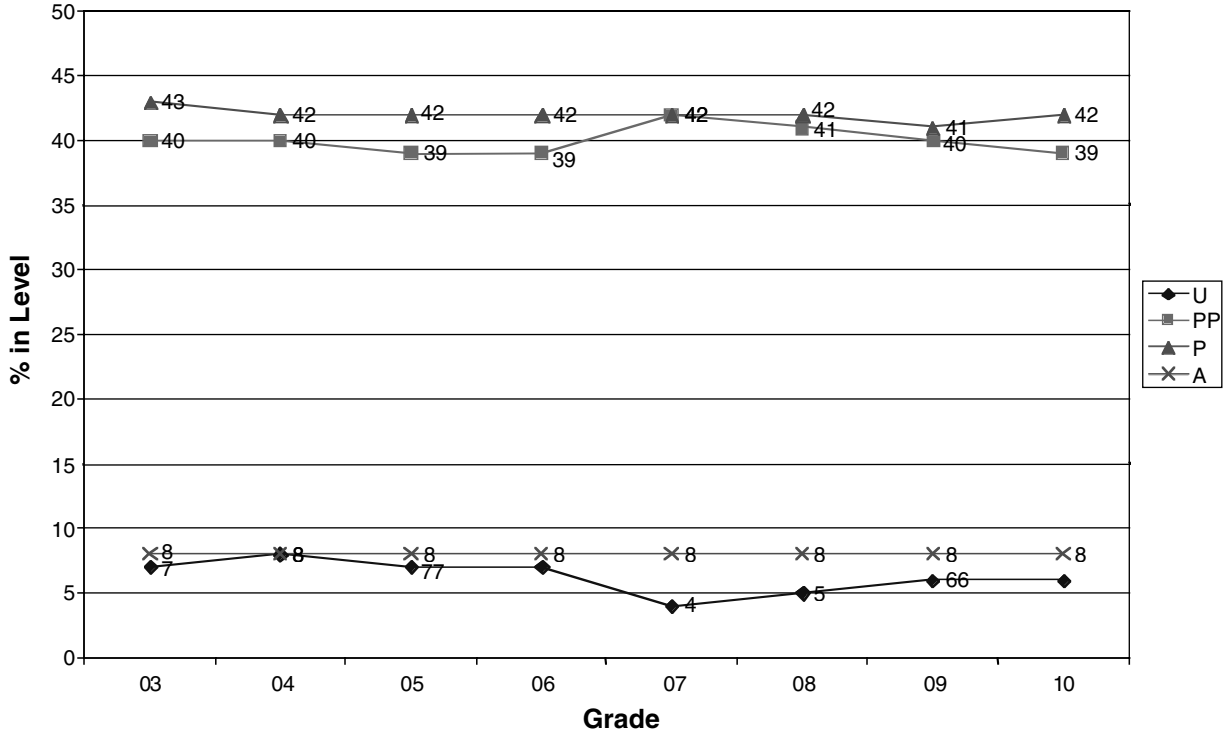
CO Math 2002



CO Math 2005



CO Writing 2002



CO Writing 2006