



Consistency for State Achievement Standards Under NCLB

Second in the Series

June 2005

**by
Howard C. Mitzel, Pacific Metrics**

**with
CAS SCASS Study Group: Transitions in
Assessment from IASA to NCLB**

Acknowledgments

This paper resulted from the work of the Study Group on Transitions in Assessments from IASA to NCLB comprised of state educational specialists and consultants of the Comprehensive Assessment Systems for ESEA Title I (CAS) State Collaborative on Assessment and Student Standards (SCASS). The members of the Study Group benefited from discussions among SCASS colleagues throughout 2004 and 2005:

Mildred Bazemore, North Carolina
Dale Carlson, Consultant
Tim Crockett, Measured Progress
Carol Crothers, Nevada (Co-chair)
William J. Erpenbach, Consultant
Arthur Halbrook, CCSSO
Tammy Howard, North Carolina
Susan Ketchum, Wisconsin
Bernadette Morris, Louisiana
Les Morse, Alaska
Jason Nicholas, Wyoming
Ed Roeber, Michigan
Grace Ross, ED, Ex-Officio
Cheryl Schroeder, Wyoming
Alan Sheinker, CTB
Gary Skoglund, South Dakota (Co-chair)
Rodney Watson, Louisiana
Charles Wayne, Pennsylvania

Jan Sheinker, CAS SCASS Coordinator

This paper was supported entirely by funding from member states of the Comprehensive Assessment Systems for ESEA Title I State Collaborative on Assessment and Student Standards (CAS SCASS), through the Council of Chief State School Officers. Information about the CAS SCASS is available on the CCSSO website, <http://www.ccsso.org>.

This publication and any comments, observations, recommendations, or conclusions contained herein reflect the work of the authors. They do not necessarily reflect the views of the Council of Chief State School Officers.

Copyright © 2005 by the Council of Chief State School Officers. All rights reserved.

Abstract

The advent of the No Child Left Behind (NCLB) legislation requires testing at the contiguous grade levels 3-8, and one grade in the high school span. The accountability provisions of the legislation further require states to classify their students according to achievement level standards and measure their growth yearly in order to demonstrate compliance to adequate yearly progress goals. This implies that performance standards which are set for assessment systems at contiguous grade levels must demonstrate a new and higher level of consistency than in the past, in terms of student classifications. This paper provides suggestions and recommendations to state departments and state policymakers for implementing achievement levels at the additional grade levels. An argument is made that the new requirement for consistency is sufficiently important to be internalized as an institutional goal, even at the cost of overriding some traditional, standard-setting practices and principles such as the independence of panels at different grade levels. Further remarks are offered with regard to consistency between content areas and the requirements for assessments to perform in the manner intended from one year to the next.

Introduction and Background

Implementation of the No Child Left Behind (NCLB) legislation has required states to test at many additional grade levels and, correspondingly, to establish achievement levels for each additional grade level tested. Under the previous IASA, most states administered their criterion-referenced tests and set achievement standards at several, usually separated, grade levels (e.g., 4, 8 and 10). Under NCLB, all states must test in reading and mathematics at grades 3 through 8 and one grade in the high school span. In 2007-08, science assessments are to be given at three grade levels.

In addition to implementing new tests for their students, the task before virtually all 50 states is now to extend their achievement levels across the added grade levels. Achievement levels are implemented through cut scores on test scales, and the stakes are high for U.S. education and educators. The adequate yearly progress (AYP) provision of the NCLB law prescribes sanctions that depend on annual measurable objectives (AMOs) and/or intermediate goals (IGs) which include consistently increasing the percentage of students placing in the proficient achievement level or above. Although there are several ways to compute AYP (i.e., grade averaging, separate computations for each grade level), it is important to recognize that under NCLB's school level accountability requirements, the entire school is the unit of accountability. Especially when grade averaging occurs, it follows that there will be strong expectations for consistency on the part of parents, educators, and students in the year-to-year achievement of each grade cohort. Achievement standards instantiate expectations for students, and these expectations should be consistent or "coherent" (Lewis & Haug, 2003) across the educational system. 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.

Continuing public acceptance of a state assessment system, as well as the accountability system, will be heavily dependent on the perceived consistency of the yearly outcomes as a grade

cohort moves through the public education system. The term “consistency” is meant to imply a special case of stability. Consistency, as used here, refers to explainable or even rational levels in (or changes in) achievement levels regardless of progress in achievement. In recent years, state policymakers have been properly hesitant to modify the recommendations of their standard-setting committees. Cut scores have tended to be adjusted cautiously within statistical error bands prescribed by psychometricians. Under the new federal accountability provisions prescribed for contiguous grade levels, state policymakers may need to trade off some traditional conventions around the acceptance of cut scores to obtain consistency across grade levels. Similarly, state departments faced with setting standards at additional NCLB grade levels should consider new designs for standard setting workshops that may require some compromise with traditional practices. In short, consistency of standards should be internalized by state departments as one of their goals (Kane, 2001) in the adoption of new achievement standards.

The focus of this paper is to identify and discuss those points at which state departments can exert some leverage to instill greater consistency into achievement levels. Of course, standard setting is one of the most important leverage points. Next, three scenarios are identified that differentiate the positions state departments may take with the setting of new standards. These scenarios are not intended to represent choices under which states may adopt any of several options but rather prototypical situations which characterize the legacy of existing achievement standards. Some states may find that more than one scenario characterizes their current situation. Later, a brief section addresses consistency between content domains (e.g., reading and mathematics), and elements that affect year-to-year consistency (e.g., form construction and equating practices) are identified.

Scenarios for Setting Achievement Level Standards under NCLB

Scenario 1: Extending Existing Achievement Standards to Additional Grade Levels

We believe this is the most common scenario among the states and the one to which we will devote the most attention. By “extending existing achievement standards” there is a presumption that satisfactory achievement standards based on criterion-referenced exams already exist at intermittent grade levels. Consider Table 1 which shows hypothetical percents of students in four achievement levels for existing assessments at grades 4, 8, and 10.

Two observations are relevant here. First, standards appear to be lower at the elementary relative to the middle and high school levels, based on percents of students falling at the proficient (P) and advanced (A) levels in Table 1. This apparent disparity between the elementary and higher grade levels is, in fact, quite common in state standards. In the author’s experience, the rationales offered by standard-setting panelists tend to center around not wanting to early “label” students as failing, but wanting to identify those that are clearly in need of remedial work. There is also a commonly articulated desire to “send a message” to school staff that instruction will need to improve for students to meet future standards.

Table 1: Hypothetical Percents of Students in Four Achievement Levels

Grade	Reading				Mathematics			
	BB	B	P	A	BB	B	P	A
4	8	43	41	8	10	44	41	5
8	14	48	33	5	22	48	28	2
10	16	47	34	3	23	47	29	1

A second observation pertains to the apparently higher standards imposed in mathematics relative to reading at the middle and high school levels. Larger percents of students fall into the below basic (BB) and basic (B) levels in mathematics relative to reading at grades 8 and 10. Again, this is not uncommon for both mathematics and science standards relative to reading/English language arts and social studies. The discourse at standard settings often centers on the perception that mathematics (and science) instruction is in need of a great deal of improvement. From this point of view, it is not that mathematics and science standards are higher relative to the other subject areas, but that students have more to learn. Again, we make the point that standards should represent the expectations of the educational community for student achievement and not just the percents of students in each achievement level. In passing, we also note that skill in mathematics can be assessed in both classroom and standardized testing situations more directly and discretely relative to reading. A related point is that the problem-solving orientation of mathematics instruction is more closely attuned to the mode of mathematics assessment. Furthermore, at the middle and high school levels, teachers are specialized for mathematics instruction. These factors may combine to make a significant contribution to the confidence and consistency with which we often see mathematics standard setting committees making their recommendations.

Under this scenario, the existing standards are viewed by the relevant policymakers as satisfactorily consistent across the existing grade levels to form the basis for setting standards at the additional grade levels. The differences between the grade 4 and 8 results that were created for Table 1 are probably close to the outside boundary for consistency.¹ Given that accountability models turn on the percents of students at or above proficiency, judgments of the consistency of standards across existing grade levels are best made at these aggregated levels. Table 2 reports the percents of students at or above the proficient level using the percents shown in Table 1.

Table 2: Percents of Students at or above Proficient Based on Table 1

Grade	Reading	Mathematics
4	49	46
8	38	30
10	37	30

Mathematics shows the largest discrepancy between grades 4 and 8. Clearly, any state that must move an additional 16 percent of its students into the proficient level, in addition to the yearly

¹ See also Lewis and Haug (2003) for additional examples of consistency vs. inconsistency in achievement levels.

AYP goals, has a considerable task ahead, but this is consistent with the realities that exist in many states. The desirable outcome, of course, is to set cut scores at the three added grade levels (5, 6, and 7) that graduate or spread this additional 16 percent of students in mathematics and 11 percent in reading. Some psychometricians refer to this process as “smoothing,” a term probably derived from norming work. Lissitz and Huhyn (2004) refer to the outcome of this smoothing process as “vertically moderated standards,” viewing the process as a type of equating procedure as discussed by Mislevy (1992).

Table 3 shows one potential solution for the hypothetical state example presented in Tables 1 and 2 for the added grade levels. These percents represent **targets** for the sum of the percents of proficient and advanced students, while the complement represents overall targets for the sum of the basic and below basic achievement levels. A second step would establish targets, through a simple interpolation process, for each of the four achievement levels at each additional grade level and content area. Note that in Table 3, slightly higher expectations are established for the first additional test administration, at grade 5. The target cut scores are derived by backing out the percents to a scale score from a frequency distribution. This procedure should be done on operational census data or very good operational samples. Field test data often tend to be too unreliable for this process.

Table 3: Vertically Moderated Percents of Students at or above Proficient Based on Table 1

Grade	Reading	Mathematics
4*	49	46
5	45	41
6	42	37
7	40	34
8*	38	30
9	37	30
10*	37	30

*Hypothetical existing state standards

It is important to keep in mind that the percents within each category represent a baseline or starting point for students in their first year. Some fluctuation due to factors such as equating error will, of course, take place in these percentages every year regardless of the amount of real growth occurring. Some of these factors are briefly described in a later section of this paper.

Once vertically moderated targets are set in terms of percents and corresponding cut scores, there remains the practical issue of how to implement the target cut scores. One possibility, as recommended by Lissitz and Huynh (2003), is for state departments of education to recommend the new standards directly to their state boards. This procedure was apparently applied recently in two states (Huynh et al., 2004). The disadvantage here is that this type of process bypasses the usual inclusive standard setting workshop, a process intended to involve the educational community and often other stakeholders in the establishment of achievement standards. The advantage, of course, is that consistency is established with a minimum of cost and effort. This process is not consistent with the recently released peer review guidance from the U.S. Department of Education (ED, 2004)

that essentially adopts NAEP language for input and review of standards by educational stakeholders. In essence, standards must continue to be set in a “socially moderated” environment.

Standard setting under the standards-based movement in the U.S. has always involved a highly pluralistic process as instituted some years ago under policies established by NAGB for NAEP achievement level setting and then taken up by most states. At the state level, an important function of standard setting is developing buy-in from the panelists for the recommended achievement standards across the state. Next, some recommendations are offered to states that plan to convene standard setting panels to establish standards for additional grade levels.

To briefly recapitulate, under the present scenario, existing standards (i.e., cut scores) are to be retained and function as anchors at intermittent grade levels, and new standards are to be set by panels for the additional grade levels. Given that cut scores, which define the achievement levels, are to be set through standard setting workshops, the following general recommendations are offered for the design of these workshops:

1. Make cross-grade consistency in the outcome part of the charge to the panelists. In other words, panelists should be told from the outset that consistent standards across grade levels are necessary for the public to retain confidence in the assessment system. A state’s top education policy officer, such as the commissioner or secretary of education, can be quite effective in making this point as a keynote speaker. It is important to express confidence in the existing standards and to characterize the setting of standards at the additional grade levels as an extension of the earlier effort. This is an opportunity to remind panelists that in most states they can only recommend the standards that are actually set by state policy bodies.
2. As a corollary to #1 above, adopt general policy descriptions, by content area and achievement level, which are the same across grade levels (Schafer, 2004). NAGB along with many states have followed this practice since the outset of the standards movement.
3. Specific achievement level descriptors can be adopted or refined following the achievement level setting process by examining the content that students must be able to do to perform at a given level. We note this is not consistent with NAGB policies, where achievement level descriptors are created prior to the achievement level setting workshop.
4. Consider the placement of provisional impact data. Provisional impact data are feedback to panelists of the percents of students in each achievement level, given cut scores do not move from their present position later in the standard setting process. Traditionally, feedback data are presented prior to the last round in standard setting as a final check on the reasonableness of the provisional cut scores. The goal of consistency would imply that feedback is shown earlier in the process. The challenge, of course, for the design of the

process is to prevent panelists from using that information in a purely norm-referenced or “empirical” manner.

5. Compose panels with cross-grade membership. At each panel or table include participants (i.e., teachers) 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. This is also an opportunity for state departments of education to review and revise achievement level descriptors at the existing grade levels.
6. Depending upon the specific standard setting method, modify the procedure in a way that the target cut scores are apparent to the panelists and their consideration becomes a part of the process. Although some panelists will likely voice objections, where targets do exist, they should be shown to panelists and openly discussed. Note that this represents yet another compromise, in this case to the traditional independence of the panelists.

With regard to the fourth point above, the new procedure piloted by ACT for NAEP grade 12 mathematics alternates a traditional bookmark approach with percent correct performance data akin to that used in the item rating methods (see below). However, impact data in terms of the (projected) percent of students in each achievement level are still held until just prior to the final round.² Suggestions specific to several general categories of standard setting methods are briefly presented next.

Item Mapping Methods. We begin with the bookmark procedure (Mitzel, Lewis, Patz, & Green, 2001; Lewis, Mitzel, & Green, 1996) and its item mapping variants as they are currently the most common standard setting procedure applied in K-12. Test items are ordered in a booklet from easy to difficult. Panelists provide a cut score judgment right on the test scale by placing a bookmark which divides the test content between two achievement levels. This process is ideal for placing target cut scores in front of panelists. The target cut scores can be represented to panelists with pre-placed bookmarks in the ordered item booklets or in other ways on ancillary materials. Louisiana State Department of Education staff has, in fact, applied this technique since the outset of the current assessment system in 1999. Target cut scores were based on projections, where applicable, from NAEP estimates for the State.

Student Work or Profiling Methods. These procedures present panelists with multiple examples of actual student work (e.g., Kingston, Kahl, & Sweeney, 2001) or provide score profiles (Jaeger, 1994) usually as graphic stimuli, along with student work samples. These methods tend to work best for assessments that are primarily composed of student constructed responses such as writing essays, science projects, portfolios, or, more recently, alternate assessments (Morgan, 2004). For these methods, panelists can be shown exemplars of student work that are consistent with the target cut score. In terms of an actual process, panelists can be shown the student work (and/or profiles) first that are consistent with the target cut scores. A “Goldilocks” exercise can be designed

² A presentation on this method is scheduled for the 2005 CCSSO National Annual Conference on Large-Scale Assessment.

in which panelists express their judgment as to whether the exemplar stimuli is about right, or a little above or below where a standard should to be placed.

Modified Angoff or Other Item Aggregation Methods. This class of standard setting procedures requires panelists to produce a judgment for each item for each achievement level. The process does not require the application of item response theory (IRT) and is commonly applied in certification and licensure applications, where sample sizes are inadequate for IRT application. Existing NAEP standards were set with a modified Angoff process, but NAGB recently changed to a variant of the bookmark procedure for grade 12 mathematics. In a modified Angoff process, panelists judge the probability of success on each item for a student classified at the nadir of the achievement level. These judgments are aggregated and usually averaged over panelists to produce recommended cut scores for each achievement level. Unfortunately, this class of methods does not lend itself well to providing stimuli to panelists that focus attention on the cut score targets, given the atomized process by which judgments are provided for individual items. One solution is to provide an ancillary stimulus that shows the expected p-value for each item at every possible scale score (Reckase, 2002). These ancillary tables were introduced in NAEP standard setting as a way to help panelists provide more consistent item-level judgments. By identifying a row of the chart, panelists can see an expected score for each test item. Target cut scores then can be identified by focusing panelists' attention on selected rows of the chart.

Under scenario 1, the recommendation is that cut score target levels be integrated into a procedure that emphasizes consistency as one of its outcome goals.

Scenario 2: Reestablishing Achievement Levels at Existing Grade Levels Where Inconsistent Standards Exist

Under this scenario, existing standards are deemed too disparate to apply vertical moderation methods at the additional grade levels. For example, existing standards between grades 4 and 8 may be too inconsistent to set a standard for grade 7 that will appear to be consonant with the achievement standards at the adjacent grades. A second possibility is that existing standards are judged to be too extreme, and there is a desire to adjust them. The idea, then, is that some existing achievement standards can be retained, while others need to be reset or at least reexamined. This is a difficult situation for state policymakers; the very term "standards" is meant to imply to the public an invariant unit of measure, or in this case a level of academic achievement. To the public, it may appear to be backtracking to change existing standards that policymakers once endorsed. There are, however, several circumstances under which existing standards can and should be reestablished.

Changes to state content frameworks. Often state content frameworks (i.e., state academic standards) have a lifespan of only about five years. Changes to the frameworks will often require corresponding adjustments to the test blueprints. Some states, such as Oklahoma, are also looking at their content from a vertical viewpoint, that is, checking the alignment and articulation of curriculum from one grade level to the next. This could also result in changes to the frameworks or re-emphasis of content within the frameworks.

Misalignment of test blueprints to content frameworks. Under NCLB, states are verifying the alignment of test blueprints to their state frameworks, which could result in adjustments to test blueprints.

Re-weighting of test blueprints. Some states have recently reduced the constructed-response portion of their assessments, presumably to save on hand scoring costs and to improve turnaround time for results. California's Board decided to reduce the "testing burden" on students and removed one of the two essays on the high school exit exam. New standards for the exam were set in September 2003.

If any of these circumstances do apply, then new standards should be set at least for those existing grade levels. Resetting or at least reviewing all of the achievement standards in a given content area should also be considered (see Scenario 3, below).

In Delaware, policymakers have been fielding criticism for several years that achievement standards are too high in mathematics for grades 8 and 10.³ There is a perception among many district level staff in Delaware that the mathematics cut scores represent unrealistic expectations for students in those two grades. In addition, a new test form is coming online for 2004-05. Following some public discussion, the current plan is to review and revisit all of the mathematics and reading achievement standards at all relevant grade levels in the summer of 2005, and establish new achievement standards at the additional grades. This brings up the third scenario, discussed next.

Scenario 3: Establishing New Achievement Standards at Contiguous Grade Levels

Here the presumption is that a clean slate exists to set standards at all applicable grade levels. A process that incorporates stakeholders from the educational community is recommended, for all of the standing values around inclusion and stakeholder buy-in. That said, a clean slate does not preclude the design of a standard setting process that incorporates cut score targets as described under Scenario 1, although this will be a methodological deviation for many states. As noted earlier, however, Louisiana has applied target levels based on NAEP since the outset of its current assessment system. Targets could be justified under previous standards, NAEP results, or some other external test results such as an NRT. Most of the same general recommendations made under Scenario 1 also apply here:

1. It is recommended to try to set all of the applicable grade levels, at least for a given content area, at the same time. This will provide the opportunity for separate panels to meet and reconsider inconsistent standards. Alternatively, if standards are set sequentially (e.g., elementary, middle, and high school), subsequent panels may object on the basis that the emphasis on consistency has predetermined the outcome.

³ Personal communication, Valerie Woodruff, Secretary of Education, Delaware Department of Education, October 2004.

2. Make consistency a charge of the standard setting process and build opportunities to resolve inconsistencies into the procedure.
3. 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.

As alluded to above, designers of a standard setting workshop that emphasize a goal of grade-to-grade consistency will want to take into account some well-known behaviors of standard setting panels. For example, the frequent (though not universal) reluctance of panelists to readjust their cut scores following the second round should be taken into account at the design stage. In the past, where standards have usually been recommended at intermittent grade levels, the practice was to keep the panels and their intermediate outcomes independent from one another to avoid undue cross-contamination. Final results to be recommended to state policy making bodies were usually only shared at the end of the process as instituted in the NAEP procedures (National Assessment Governing Board, 1995). The recommendation here is that this traditional practice of panel independence should be re-examined and probably subjugated to the priority for cross-grade consistency of achievement standards.

On the other hand, some panelists and state department staff may believe that apparently inconsistent standards, as indicated by grade-to-grade disparities in the percentages of students in achievement levels, are in fact correct and proper. This position may be based on evidence that instruction or some other condition unique to the state has influenced student achievement. Misalignment among state academic standards, instructional practices, and/or corresponding test blueprints would be a likely cause. For example, suppose content frameworks and test blueprints were recently revised for writing at grades 4, 5, and 6, but there is adequate evidence that instructional practices in the state have not caught up to the change. In this case, student achievement will suffer due to the lack of alignment, and establishing higher initial percentages of students below the proficiency level at any additional grade levels would seem justified. The alternative would be to reward complacency in modifying instructional practice by setting relatively easier standards at those grade levels.

Clearly these types of situations are difficult for state decision makers and should be based on some tangible evidence such as worse-than-expected student performance or survey data pointing to instructional misalignment. In the author's experience, these types of instructional misalignment hypotheses can be advanced during a standard setting as an explanation for data showing poor student performance, especially in science and mathematics. Whether or not there is actual evidence to support it, the hypothesis can become a shared belief among panelists that may affect the level at which cut scores are set. These are not the type of surprises that are welcome to state department personnel, and standard setting workshops may need to be monitored more closely than in past years.

Consistency between Content Areas

Louis Guttman often stated that if universal laws were to be established for scientific psychology, the first law of psychology should be that all human performance is positively correlated.⁴ Our expectation, both as educators and members of the public, is that performance in one content domain should be similar to performance in another given a common group of students. The problem here concerns how closely percentages of students in achievement levels within a single grade level should mirror one another, and how much adjustment can be tolerated to achieve consistency. Lissitz and Huynh (2003) refer to this outcome as horizontally moderated standards. The goal for policymakers, then, becomes to establish consistency both within and across content areas. Years of achievement testing have shown mathematics and science are usually more closely correlated to one another than they are to reading, writing, and social studies, and policymakers may use this as justification to adjust some cut scores.

One caution is offered related to earlier comments around across-grade consistency. There may be, in fact, relative differences in student achievement between content domains that would justify apparently disparate percents of students in achievement levels. Again, such disparities will likely be due to alignment issues or statewide instructional practices. The author has worked in one state in which department staff believe mathematics instruction is generally deficient at the middle and high school levels, and have indicated that “reasonable” standards will show far more students below proficiency in mathematics than in reading. NAEP results are expected to confirm their beliefs. The point here is essentially similar to that made above for vertical standards; there may be situations in which aligning standards by purely statistical means will result in a disservice to students and the rest of the educational community. These situations need to be identified in advance of standard setting or policy level adjustments to test cut scores.

In general, there is probably more public tolerance for apparent disparities in standards horizontally as opposed to vertically. Where horizontal disparities among standards are endorsed through policy decisions, it is important to explain to the public where and why these discrepancies exist. Again, external data from NAEP or some other source will be important in validating the policy.

Other Factors in Year-to-Year Achievement Consistency

Thus far, this paper has been primarily concerned with consistency in student performance across grade levels. We now turn to the problem of consistency for a large-scale assessment **within** a grade level and content domain across time. From one year to the next, academic growth will be one and hopefully the largest factor affecting change in the percents of students in each achievement level. Inevitably, other factors, which essentially amount to noise in the system, will contribute variation and resulting uncertainty, in the year-to-year results. Many of the alignment issues discussed earlier also apply in this situation, if there is change over time. A technical challenge

⁴ Personal communication, Louis Guttman, University of Chicago, July 1983.

before state departments is to minimize the effect of these nuisance factors on the assessment system. In this section, we provide only a general overview of this topic by identifying some of these nuisance factors.

The number and nature of achievement levels. Given that we have identified consistency as a new priority in establishing achievement standards, there are perhaps few mechanisms more ill-suited to reporting those standards than percents in categories. The problem is that these percents are dependent on one another. As the percent in one category changes, it will affect one or more of the other achievement levels. A simple recommendation is to create student accountability systems with no more than four achievement levels. Many states have five levels, which include subdivisions of levels below proficiency for state accountability purposes. This is likely to become problematic over time due to test reliability in the lower achievement ranges for most state tests. In the public reporting of assessment results, states must report in all of the achievement levels. States are urged to commission simulation studies that can be customized for each state's assessment design which can model the stability of their achievement levels (and corresponding accountability systems) over time given, of course, certain limiting assumptions. One possible finding from these studies is that some achievement levels may be so unreliable as to make reporting in them essentially useless in terms of the original goals of reporting.

It is untrue that classification error "averages out." More achievement levels in fact result in more classification error. This will translate to instability in the percents of students at each achievement level, depending on several factors. Factors affecting classification error include (1) how the reliability characteristics (i.e., the conditional standard error of measurement) vary across the test scale, (2) density of students around the cut scores, (3) number and width (in terms of scale score units) of the achievement level, and (4) a near myriad of potential issues related to the measuring properties and construction practices of the assessment instrument. Some of these are noted next.

Measuring properties of the assessment instruments. As large-scale testing has moved from predominantly norm-referenced to criterion-referenced applications over the past decade, testing administrators have become increasingly familiar with requirements for test validity and reliability. As many states have supervised the construction of their own assessments over the past decade, many state technical advisory committees (TACs) have been formed, and state department staff have become increasingly sophisticated in the technical issues of test validity and reliability. Longer tests (i.e., tests with more items) of course have better reliability than shorter tests. This translates to fewer errors of misclassification and will result in better stability for year-to-year estimates of the percents of students in achievement levels. If blueprint and test construction issues are still on the table, states and their TACs should look carefully at supplementing tests in the scale score ranges where important cut scores occur. Using modern item response theory (IRT) tools, it is now possible to target appropriate content at student achievement in terms of the test scale. In other words, it is possible to increase the reliability of a test in the region where it is needed the most.

Development, construction, and technical practices. In some states, assessments used for NCLB overlap with the tests used for student promotion or graduation requirements. In these cases, where high stakes decisions depend at least partly on test scores, new forms are constructed each

year with embedded or common item designs needed for linking (i.e., equating) from one administration to the next. This means, for assessment systems which operate under these designs, a large proportion of the items in a test form are changed out with each administration. It is nearly inevitable that even when test blueprints are carefully adhered to, over time, these item substitutions will affect how and what a test measures differentially across the test scale. This effect will be exacerbated over time by changes in item writers and even development contractors. Technically, this will result in some unevenness in the scales of forms that are intended to be equivalent from one administration to the next, in turn, affecting the articulation of the test scales under equating. For states utilizing augmented NRT forms, this effect is probably mitigated to some degree, depending on the proportion of the test that is supplemented. However, there is always some effect from item exposure occurring when items are repeated year after year, which can lead to what is usually referred to as scale drift (Mitzel, Weber, & Sykes, 1999).

Many of the steps that can be taken to minimize these effects should occur during item development and form construction stages. When test content is exchanged to create a new form, items or groups of items (e.g., strands) that measure a given objective or construct should be at the same difficulty level from one form to the next. This goal needs to be stated in the test design specifications, and will likely increase the cost of form development. In essence, this requires some type of form pre-equating, as simple post equating processes will not address nonlinearities potentially introduced into the measurement scales. For example, if an eighth grade test blueprint were to specify that each form contain three items that measure knowledge of the Pythagorean Theorem, then that strand should ideally operate at the same level of difficulty in each form. Most developers that do perform technical pre-equating studies only assure the form as a whole is pre-equated, not the components of the form.

Depending on the specific linking design for a test form, it is often not recognized that the degree of equating error associated with sampling fluctuations can be significant. For an “average” state test with a standard deviation of 50, equating error on the order of several scale score points can be within a normal range. This can easily shift the magnitude of students in a given achievement level by 5 percentage points near the densest part of the score distribution. Again, this is not the type of error that is somehow “averaged out” on the other side of a cut score. Larger equating samples can be used to mitigate equating error, and TACs and state department staff should be looking at their assessment practices with a fresh eye to improving their technical quality.

Conclusion

One outcome of the standards movement is that it has replaced the focus on scale scores with a focus on performance labels (e.g., basic, proficient, and advanced). It seems these labels are easily reified by the public such that achievement levels instantiate our expectations of student performance on modern summative assessments. As NCLB requires testing and accountability at contiguous grade levels, grade-to-grade consistency in percents of students within these achievement levels will become crucial for continued public support of state implementation of adequate yearly progress and state accountability systems. Suggestions for “vertically moderated standards” are essentially suggestions to arithmetically interpolate between existing achievement

standards. Federal peer review guidelines, however, continue to require traditional review and approval processes involving the diversity of educational stakeholders (e.g., teachers, content specialists, special education experts, minorities, members of the general public), such that achievement levels defined through entirely empirical methods are not likely to be approved. One suggestion here is to convene standard setting workshops that utilize the vertically moderated standards as target values for the workshops.

Many of the modern standard setting techniques were designed with the goals of standards-based assessment practices in mind. The bookmark procedure, for example, orders test content according to student achievement and elicits a direct judgment of a cut score from panelists. In this way, panelists are able to see directly what content must be mastered to gain access to an achievement level. Similar goals exist for standard setting methods that examine and classify student work. By directly examining student work, we can determine what a student should know and be able to do. It is an irony, then, that in attempting to establish new (vertical) achievement standards under NCLB that compromises and even methodological backtracking may occur in the methods so carefully developed for a standards-based educational system. States need to begin a discussion with their educational communities as to how vertical standards should be set, and what compromises we are willing to tolerate and still declare that the establishment of new standards has followed a valid process (see Hambleton, 2001).

It is already late to begin a national discussion of the application of vertically-moderated standards. We should be clear that the use of target achievement levels for additional grades, as suggested here, has much the same role as beginning a standard setting with impact data, although some nuance exists in the method of presentation of the targets to panelists. Our discussion of techniques like this should resolve to our satisfaction if we are, in fact, abandoning the application of content-based standards and instead substituting normatively-based standards.

As we consider how to set new achievement standards, it is perhaps time to take a fresh look at the technical quality of our state assessments. There has been considerable recent attention paid to the horizontal alignment of instruction to assessment. The vertical alignment of content also seems crucial for instruction to articulate with assessment from one year to the next. Overlapping academic standards will only create conflict, confusion, and waste in the limited space that assessment occupies. As student cohorts move through public instruction, classification error and other nuisance factors will obfuscate their true growth rates. Larger samples may be applied to reduce calibration and equating error. Most of the year-to-year assessment forms in use are not parallel in the strong sense of the term, and this will result in nonsystematic errors in their linking. State departments and TACs should perhaps reexamine contractors' test development and construction practices to see if more scientific approaches can be utilized. In short, state departments and their TACs should look to every opportunity to ensure that future growth estimates are as uncontaminated as possible.

References

- Hambleton, R. K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In Cizek, G. J., (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 89-116). Mahwah, NJ: Lawrence Erlbaum.
- Huynh, H., Barton, K.E., Meyer, J.P., Porchea, S., & Gallant, D. (2004, June). *Vertically moderated standards for SC PACT 1999 assessments of English language arts and mathematics: A look back from adjacent-grade student data*. Paper presented at the meeting of the Council of Chief State School Officers, Boston, MA.
- Jaeger, R. M. (1994, April). *Setting performance standards through two-stage judgmental policy capturing*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.
- Kingston, N.M., Kahl, S.R., Sweeney, K.P., & Bay, L. (2001). Setting performance standards using the body of work method. . In Cizek, G. J., (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 219-248). Mahwah, NJ: Lawrence Erlbaum.
- Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting standards. In Cizek, G. J., (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 249- 281). Mahwah, NJ: Lawrence Erlbaum.
- Lewis, D. M. & Haug, C.A. (2003). *Aligning policy and methodology to achieve consistent across-grade performance standards*. Monterey, CA: CTB/McGraw-Hill.
- Lewis, D. M., Mitzel, H. C. & Green, D. R. (1996, June). Standard setting: A bookmark approach. In D. R. Green (Chair), *IRT-based standard-setting procedures utilizing behavioral anchoring*. Symposium conducted at the Council of Chief State School Officers Annual Conference on Large-Scale Assessment, Phoenix, AZ.
- Lissitz, R.W. & Huynh, H. (2003, January). *Vertical equating for the Arkansas ACTAAP Assessments: Issues and solutions in determination of adequate yearly progress and school accountability*. Report submitted to the Arkansas Department of Education, Little Rock, AR.
- Mislevy, R. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects*. Princeton, NJ: Educational Testing Service, Policy Information Center.
- Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The bookmark procedure: Psychological perspectives. In Cizek, G. J., (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 249-281). Mahwah, NJ: Lawrence Erlbaum.
- Mitzel, H. C., Weber, M. M., & Sykes R. C. (1999, April). *Test item disclosure: How much difference does it really make?* Paper presented at the meeting of the National Council on Measurement in Education, Montreal, Canada.

- Morgan, D. (2004, June). *The performance profile method: a unique standard setting method as applied to a unique population*. Paper presented at the Council of Chief State School Officers Annual National Conference on Large-Scale Assessment, Boston, MA.
- National Assessment Governing Board. (1995). *Developing student performance levels on the National Assessment of Educational Progress*, (amended March 4, 1995). See Appendix A.
- Reckase, M.D. (2002, April). *Improving panelists' understanding of the standard setting process using item response theory*. Paper presented at the meeting of the National Council on Measurement in Education, New Orleans, LA.
- Schafer, W.D. (2004, April). *Standard setting from a state perspective*. Paper presented at the meeting of the Council of Chief State School Officers, Boston, MA.
- 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: U.S. Department of Education, Office of Elementary and Secondary Education.