Listening. Learning. Leading.®

R&D Connections
No. 7 • October 2008

Standard Setting: What Is It? Why Is It Important?

By Isaac I. Bejar

Standard setting is a critical part of educational, licensing, and certification testing. But outside of the cadre of practitioners, this aspect of test development is not well understood.

Standard setting is the methodology used to define *levels* of achievement or proficiency and the *cutscores* corresponding to those levels. A cutscore is simply the score that serves to classify the students whose score is below the cutscore into one level and the students whose score is at or above the cutscore into the next and higher level.

Clearly, unless the cutscores are appropriately set, the results of the assessment could come into question. For that reason, standard setting is a critical component of the test development process.

Unless the cutscores are appropriately set, the results of the assessment could come into question.

This brief article does not address the technicalities of the process, for which readers can consult several references (Cizek & Bunch, 2007; Hambleton & Pitoniak, 2006; Zieky, Perie, & Livingston, 2008). Instead, this article illustrates the importance of standard setting with reference to accountability testing in K-12 and suggests that some of the questions that have emerged concerning standard setting in that context can be addressed by considering standard setting as an integral aspect of the test development process, which has not been standard practice in the past.

In tests used for certification and licensing purposes, test takers are typically classified into two categories: those who "pass"—that is, those who score at or above the cutscore—and those who "fail."

These types of tests, therefore, require a single cutscore. In tests of educational progress, such as those required under the No Child Left Behind Act (NCLB), students are typically classified into one of three or four achievement levels, such as *below basic, basic, proficient,* and *advanced* (United States Congress, 2001). As a result, with four achievement levels, three cutscores need to be determined.[1]

In a K-12 context, decisions based on cutscores affect not only individual students, but also the educational system.

In the latter case, group test results are summarized at the school, district, or state level to determine the proportion of students in each proficiency category. As part of NCLB legislation, for example, a school's progress toward educational goals is expressed as the proportion of students classified as proficient.

So, how do we know if the cutscores for a given assessment are set appropriately? The

[1] NCLB is an example of standards-based reform. It differs significantly from previous attempts at educational reform characterized by "minimum competency." According to Linn and Gronlund (2000) standards-based reform is characterized by the adoption of ambitious educational goals; the use of forms of assessment that emphasize extended responses, rather than only multiple-choice testing; making schools accountable for student achievement; and, finally, including *all* students in the assessment.

"right" cutscores should be both consistent with the intended educational policy *and* psychometrically sound.

The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & American Council on Measurement in Education, 1999) suggest several soundness criteria, such as: "When proposed score interpretations involve one or more cutscores, the rationale and procedures used for establishing cutscores should be clearly documented" (p. 59). The accompanying comment further states that "Adequate precision in regions of score scales where cut points are established is prerequisite to reliable classification of examinees into categories." (p. 59).

Differing State Policies

A further criterion in judging the meaning of the different classifications, especially the designation of proficient, involves an *audit* or comparison with an external test (Koretz, 2006). Two recent reports (Braun & Qian, 2007; Cronin, Dahlin, Adkins, & Kingsbury, 2007) took that approach by examining proficiency levels across states against a national benchmark. Both studies found that states differed markedly in the proportion of students designated as proficient.

Is one state's educational system really that much better than the other? It is difficult to say by simply looking at the proportions of students classified as proficient because each state is free to design its own test and arrive at its own definition of proficient through its own standard-setting process.

However, by comparing the results of each state against a common, related, nationwide assessment, it is possible to judge whether the variability in states' proportions of proficient students is due to some states having better or worse educational systems rather than being due to the states inadvertently applying different standards.

The study by Braun and Qian (2007) used the National Assessment of Educational Progress (NAEP[2]) as the common yardstick for comparing states' proportions of students classified into the different levels of reading and mathematics proficiency against the NAEP results for each state.

NAEP covers reading and mathematics, just as all states do with their NCLB tests, but NAEP has its own definition of proficiency levels and its own approach to assessing reading and mathematics, which differs from each state's own approach. For example, NAEP includes a significant portion of items requiring constructed responses—that is, test questions that require test takers to supply their own answers, such as essays or fill-in-the-blank answers, rather than choosing from standard multiple-choice options.

Nevertheless, NAEP provides as close as we can get to a common yardstick by virtue of the fact that a representative sample of students from each state participates in the NAEP assessment.

The conclusion in the Braun and Qian (2007) and Cronin et al. (2007) reports was that the differences in the levels of achievement across states seemed to be a function of each state's definition of proficiency—that is, the specific cutscores they each used to define achievement levels. The differences in levels of achievement were not necessarily due to variability in the quality of educational systems from state to state.

In short, standard setting matters: It is not simply a methodological procedure but rather an opportunity to incorporate educational policy into a state's assessment system. Ideally, the standard-setting process elicits educational policy and incorporates it into the test development process to ensure that the cutscores that a test eventually produces not only reflect a state's policy but also are well-supported psychometrically.

[2] http://nces.ed.gov/nationsreportcard/

Cutscores that do not represent intended policy or do not yield reliable classifications of students can have significant repercussions for students and their families; fallible student-level classifications can provide an inaccurate sense of an educational system's quality and the progress it is making towards educating its students.

Setting Standards

As mentioned earlier, the standard setting process has been well documented in several sources (Cizek & Bunch, 2007; Hambleton & Pitoniak, 2006; Zieky et al., 2008). In this section, we emphasize the relationship of the standard setting process to test development.

While setting standards appropriately is critical to making sound student- and policy-level decisions, it is equally important that the content of the test and its difficulty level be appropriate for the decisions to be made based on the test results. We cannot expect a test that does not cover the appropriate content or is not at the appropriate level of difficulty to lead to appropriate decisions—regardless of how the process of setting cutscores is carried out.

We cannot expect a test that does not cover the appropriate content or is not at the appropriate level of difficulty to lead to appropriate decisions—regardless of how the process of setting cutscores is carried out.

Producing a test that targets content and difficulty toward the decisions to be made requires that item writers have a strong working understanding of those decisions. When developers design a test in this fashion, it is more likely that the cutscores will lead to meaningful and psychometrically sound categorizations of students.

This means, however, that standard setting must be done *in concert* with the test development process and not be treated as a last or separate step independent of the process (Bejar, Braun, & Tannenbaum, 2007). In fact, Cizek and Bunch (2007, p. 247) proposed that "standard setting be made an integral part of planning for test development."

The integration of standard setting into the test development process becomes more crucial in light of NCLB.[3] As part of NCLB legislation, schools test *adjacent* grades every year. Because the legislation calls for all students to reach the level of proficient by 2014, inferences about the proportion of students in different achievement categories in adjacent grades, or in the same grade in subsequent years, are inevitable because they are *prima facie* evidence about the progress, or lack of progress, the educational system is making towards the 2014 goal. More likely than not, there will be variability in the rates of proficiency in adjacent grades.

For example, one explanation for the variability in observed achievement levels across grades is that the standards across grades are not *comparable*. The cutscores that define a proficient student in two adjacent grades could, inadvertently, not be equally demanding.

This can occur if the standard-setting process for each grade is done in isolation without taking the opportunity to align the results across grades (see Perie, 2006, for an approach to the problem.) Similarly, failure to make the scores themselves comparable across years could generate variability in the proportion of students classified as proficient (Fitzpatrick, 2008).

[3] In light of the upcoming national elections in the United States, it will be necessary to monitor how federal educational policy will evolve, but there is reason to believe standard setting will continue to be part of the American educational landscape (Ryan & Shepard, 2008, p. xii).

An alternative explanation for the variability in the different rates of achievement across grades is that higher proportions of students classified as proficient do, in fact, accurately reflect a better or improving quality of education in some grades. To reach that conclusion, it is necessary to rule out the first explanation.

However, compensating for incomparable standards is a complex and unfamiliar process. In contrast, the process of equating—making scores from different forms comparable (Holland & Dorans, 2006)—has a long history.

The process of compensating for differing standards is unfamiliar because psychometricians in the United States had not dealt with the issue prior to NCLB legislation, which introduced the testing of students in adjacent grades. Since that time, the field of educational measurement has recognized the issue and proposed solutions (e.g., Lissitz & Huyhn, 2003).

However, attempting to compensate for incomparable standards *after the fact*—that is, as a prior step to releasing results—risks the possibility that satisfactory compensation may not be feasible. For that reason, it would be preferable to develop the assessments for different grades with comparable standards across grades as an explicit criterion from the start to avoid the problem as much as possible.

The foregoing complexities have motivated the formulation of alternative accountability models, as has a general dissatisfaction with the "status model" approach to accountability promoted by NCLB (Linn, 2005). Under this current accountability model, results at a single point in time are the basis for decisions as to whether a school is making adequate progress.

Several states have proposed alternative "growth models" (U.S. Department of Education, 2005). An analysis of the different features is available (Dunn & Allen, 2008). If growth models go forward, standard setting will be equally relevant to decide what growth rate is adequate (Betebenner, 2008). In short, standard setting is likely to continue to play a critical role in the future.

Conclusions

Standard setting should be seen as a critical aspect of the test development process best carried out in concert with all other aspects of the development process. Far from being a purely methodological process, standard setting ideally involves policy makers, test developers and measurement specialists early on to ensure that the test results will be useful and defensible.

References

American Educational Research Association, American Psychological Association, & American Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Bejar, I. I., Braun, H., & Tannenbaum, R. (2007). A prospective, predictive and progressive approach to standard setting. In R. W. Lissitz (Ed.), *Assessing and modeling cognitive development in school: Intellectual growth and standard setting* (pp. 1-30). Maple Grove: MN: Jam Press.

Betebenner, D. W. (2008). Toward a normative understanding of student growth. In K. E. Ryan & L. A. Shepard (Eds.), *The future of test-based educational accountability* (pp. 155-170). New York: Routledge.

Braun, H. I., & Qian, J. (2007). An enhanced method for mapping state standards onto the NAEP scale. In N. J. Dorans, M. Pommerich & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 313-338). New York: Springer.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.

Cronin, J., Dahlin, M., Adkins, D., & Kingsbury, G. G. (2007). *The proficiency illusion*. Retrieved September 20, 2008, from the Thomas B. Fordham Institute Web site: http://edexcellence.net/doc/The_Proficiency_Illusion.pdf

Dunn, J. L., & Allen, J. (2008, March). *The interaction of measurement, model, and accountability: What are the NCLB growth models measuring?* Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Fitzpatrick, A. R. (2008, March). *The impact of anchor test configuration on student proficiency rates*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Hambleton, R. K., & Pitoniak, M. (2006). *Setting performance standards*. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433-470). Westport, CT: Praeger.

Holland, P., & Dorans, N. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 187-220). Westport, CT: Praeger.

Koretz, D. (2006). Testing for accountability in K-12. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 531-578). Westport, CT: Praeger.

Linn, R. L. (2005). Conflicting demands of No Child Left Behind and state systems: Mixed messages about school performance. *Education Policy Analysis Archives, 13*. Retrieved September 20, 2008, from http://epaa.asu.edu/epaa/v13n33/

Linn, R. L., & Gronlund, N. E. (2000). *Measurement and assessment in teaching* (8th ed.). Upper Saddle River, NJ: Prentice-Hall.

Lissitz, R. W., & Huynh, H. (2003). Vertical equating for state assessments: Issues and solutions in determining adequate yearly progress and school accountability. *Practical Assessment, Research, & Evaluation*, *8*. Retrieved September 20, 2008, from http://pareonline.net/getvn.asp?v=8&n=10

Perie, M. (2006). *Convening an articulation panel after a standard setting meeting: A how-to guide*. Retrieved September 20, 2008, from the Center for Assessment Web site: http://www.nciea.org/publications/ RecommendforArticulation_MAP06.pdf

Ryan, K. E., & Shepard, L. A. (Eds.). (2008). *The future of test-based educational accountability*. New York: Routledge.

U.S. Department of Education. (2005). *Secretary Spellings announces growth model pilot study* [Press release]. Washington, DC: Author. Retrieved September 20, 2008 from http://www.ed.gov/news/pressreleases/2005/11/11182005.html

United States Congress. (2001). *No Child Left Behind Act of 2001: Conference report to accompany H.R. 1, report 107-334*. Washington, DC: Government Printing Office.

Zieky, M. J., Perie, M., & Livingston, S. (2008). *Cutscores*: *A manual for setting standards of performance on educational and occupational tests*. Available from http://www.amazon.com/Cutscores-Standards-Performance-Educational-Occupational/dp/1438250304/

Acknowledgements

I am grateful to Rick Tannenbaum, Mike Zieky, and, especially, to Dan Eignor. Their comments, I believe, have improved the article. Of course, I'm solely responsible for any remaining problems.

R&D Connections is published by

ETS Research & Development
Educational Testing Service
Rosedale Road, 19-T
Princeton, NJ 08541-0001

Send comments about this publication to the above address or via the Web at:

http://www.ets.org/research/contact.html