Bulletin
January 2012 | Issue 21
www.pearsonassessments.com

Evidence Based Standard Setting: Establishing Cut Scores by Integrating Research Evidence with Expert Content Judgments

Jennifer N. Beimers, Walter D. Way, Katie Larsen McClarty, and Julie A. Miles

Introduction

In the 1990s, a movement started that began to change the way doctors approach diagnosing and treating their patients. The movement was called *evidence-based medicine*. Sackett et al. (1996) defined the term as follows: "Evidence based medicine is the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients. The practice of evidence based medicine means integrating individual clinical expertise with the best available external clinical evidence from systematic research." Evidence-based medicine seeks to clarify those parts of medical practice that are in principle subject to scientific methods and to apply these methods to predict the best outcomes in medical treatment, even as debate continues about which outcomes are desirable.

An analogy to evidence-based medicine can be applied to the practice of setting performance standards for large-scale assessments in the United States. The most commonly used standard-setting methods focus on test content and primarily employ content experts to render judgments about the level of performance needed to achieve designated classifications such as "proficient," "advanced," or "passing." Content-centered standard-setting methods were well suited to the standards-based testing programs that proliferated under No Child Left Behind, but their application resulted in performance standards that were inconsistent across states and sometimes inadequate for their stated purposes. Too often, students labeled "proficient" on state assessments were simply not prepared for the challenges posed by college and the workplace.

In this bulletin, we describe the processes and practices associated with *Evidence Based Standard Setting*, which draw directly from the concept of evidence-based medicine. Evidence Based Standard Setting integrates content-centered judgments by appropriate experts with the best available evidence from systematic research. Therefore, Evidence Based Standard Setting requires deliberate collection and application of research data to the process of setting performance standards.

> With current education reforms focusing on preparing students for college and career, the process for developing content standards, assessments, and performance standards must shift.

College Readiness and Evidence Based Standard Setting

The U.S. education system has increasingly focused on college and career readiness. A number of individual states, as well as the two Race to the Top (RTTT) assessment consortia, aim to produce students who are prepared for college and career when they complete high school. To this end, content standards are being developed to explicitly align with definitions of college and career readiness, and then assessments are developed for use at multiple grade levels to measure student progress toward this goal. For students to be

classified into various levels of readiness in high school, standards must be set to determine the level of performance that exemplifies each readiness level. Likewise, performance standards in earlier grades must indicate the extent to which students are on track toward college and career readiness. With a shift in the desired meaning of cut scores, namely college and career readiness or being on track to reach that goal, a standard setting approach that can explicitly link cut scores to these intended uses is needed. Evidence Based Standard Setting fills this need.

From a Content-Centered to an Evidence-Based Approach

Although the most commonly used standard-setting methods focus on the content of an exam and consider the external validity of the standards well after cut scores have been determined, Evidence Based Standard Setting considers a collection of external validity data from the outset. Through this approach, various sources of data, methods of analysis, and perspectives of stakeholders can be simultaneously considered while making cut score recommendations. In addition, there can be an explicit empirical link between specific performance on an exam and what such performance indicates about likely success at the next grade level as well as progress toward college or career readiness. For example, data may indicate that students who score 53 on an Algebra II exam have a 65% chance of succeeding in their first credit-bearing college math course, or perhaps a score of 53 is empirically linked to the college readiness cut of another exam. Instead of standard setting panelists focusing solely on the content of each question on an exam, a collection of validity data pertaining to that exam could be presented for panelists to use as the basis for cut score recommendations.

In Evidence Based Standard Setting, panelists are presented with various pieces of empirical data that point toward relevant outcome variables such as success in the next grade, college and career readiness, or international competitiveness and are tasked with determining where the data most logically converge. The validity data could include a wide variety of information, such as concurrent studies, predictive studies, contrasting groups studies, and content judgment studies. Concurrent studies could be conducted to investigate the exam's relationship with related exams, such as other statewide assessments, ACT, SAT, ASVAB, NAEP, or PISA. Relevant scores on these other exams, perhaps the college readiness score on the ACT, could be empirically linked to a specific score on the exam. Predictive studies, for example, could provide success probabilities for outcomes of interest (e.g., subsequent course performance, college placement, job performance outcomes) for students earning particular scores. Contrasting groups studies could indicate the exam score that empirically separates groups of students. This could mean separating, for instance, individuals hired for particular jobs from those who struggle to find employment. Finally, content judgment studies, which could essentially be item-based standard-setting approaches, could provide perspectives from various educators and stakeholders (e.g., secondary teachers, college professors, business leaders) on the content knowledge attainment necessary for success in a class or workplace.

Requirements for Evidence Based Standard Setting

Preparation for Evidence Based Standard Setting requires a considerable amount of time and resources, making planning critical. Well ahead of the actual standard setting meeting, decisions must be made regarding what data sources are most relevant for the desired cut score inferences, how the data will be presented to panelists, and which stakeholder perspectives are most important to include in the process.

Desired cut score meanings should be the driver in determining what data are most relevant, and an efficient plan for collecting data should be implemented. In setting college readiness standards, data on existing college readiness exams or college outcomes may be useful, while data on job outcomes would be more relevant for setting career readiness

standards. Regardless of the data source, consideration should be given to how data will be collected and analyzed. To facilitate comparisons, it is particularly important to take the steps necessary to combine student performance data on a variety of measures.

Compiling validity study results and pertinent information to present to panelists will likely involve a large quantity of data. With multiple data sources being analyzed in various ways, the validity studies will be extensive, and determining which details are most relevant without overwhelming the participants in Evidence Based Standard Setting is a challenge. Synthesizing and logically organizing the data into an effective, user-friendly presentation for the panelists is crucial, because it is likely that few panelists will be familiar with the statistical methods implemented.

Though panelists for traditional content-based standard setting methods are familiar with test content and work closely with students who take the exam, Evidence Based Standard Setting encourages the use of diverse panelists. That is not to say that content experts should be removed from the process, but additional perspectives can and should be included and are often desirable. For example, college professors may be valuable to the process of setting college readiness cuts. Including business leaders may be beneficial when setting career readiness standards. In addition, when setting an aligned system of K–12 standards, having some continuity in the panelists is desirable, whether one panel makes recommendations for all grade levels or whether representatives from lower grade panels also serve on upper grade panels.

Implementations of Evidence Based Standard Setting

Recently, several entities have chosen to use validity data to set standards. The American Diploma Project (ADP) conducted a standard setting in 2009 while the Texas Education Agency plans to set standards in 2012. In both instances, variations of Evidence Based Standard Setting were, or are being, implemented.

The ADP Algebra II End-of-Course Exam was developed by the 15-state ADP Assessment Consortium to be an indicator of college readiness. Therefore, being able to link exam scores to external criteria was imperative. An Evidence Based Standard Setting approach was implemented that was loosely modeled on a "briefing book" method proposed by Haertel (2002). Panelists were presented with a briefing book that included a variety of policy background, research data, test content information, and data about student performance in a comprehensive and focused fashion designed to structure deliberations. Validity studies (concurrent studies with state- and national-level exams, cross-sectional studies, and content judgment studies) were the heart of the briefing book, but additional documents were included to provide context, explain methodology, and summarize the large quantities of information.

Data from the validity studies were analyzed with multiple methods and disaggregated by three types of post-secondary institutions and two mathematics courses. Community colleges, 4-year "typical" institutions, and 4-year "more selective" institutions[1] were sampled, and the two courses that were examined in the validity studies were College Algebra and Pre-Calculus. Panelists represented both state departments of education and higher education, providing contrasting perspectives when they considered the validity data from multiple sources.

Performance standards on the State of Texas Assessments of Academic Readiness (STAAR) will be set in 2012, informed by validity data for each grade and subject. In high school, the STAAR end-of-course assessments in Algebra II and English III will have college readiness performance standards established. Information from concurrent studies (i.e., SAT, ACT, Accuplacer, and THEA), contrasting groups studies (i.e., college students in entry-level courses taking STAAR), and content

[1] More selective institutions were determined based on admittance rates.

judgments (i.e., expert stakeholders evaluating item content) will be used to arrive at recommended college readiness performance standards.

Performance standards for other courses and grades in mathematics and English/reading will be set to indicate whether students are on track, informed by studies following cohorts of students over time. Additional studies available to inform the placement of cut scores for multiple content areas include relationships between STAAR and the previous testing program (Texas Assessment of Knowledge and Skills), course performance, and the National Assessment of Educational Progress (NAEP). Diverse stakeholders will review the results of these studies, as well as content judgments and other assessment program policies and information, to recommend performance standards. Validity evidence will continue to be gathered once the standards are in place, and the performance standards will be reviewed at least once every three years (Texas Education Agency, 2010).

How States Can Prepare

Given the significant amount of data required for Evidence Based Standard Setting, states can begin planning by developing a clear definition of the performance standards so that an appropriate data collection plan can be developed. States must develop a mechanism for obtaining the necessary student data and identifying the same students across time and multiple data sources. Additionally, state privacy laws may add complexity and should be researched before data collection.

With current education reforms focusing on preparing students for college and career, the process for developing content standards, assessments, and performance standards must shift. Evidence Based Standard Setting, which combines multiple sources of validity evidence, supports many perspectives, and provides context to the process, is fundamental to setting standards in the Next Generation assessments. While still valuing judgments of content experts, Evidence Based Standard Setting is a versatile approach that uses the best available external evidence based on systematic research to support the desired interpretations of performance standards—whether that interpretation be an inference about success at the next grade level, future success in college or career, or some other desired outcome.

References

Haertel, E. H. (2002). Standard setting as a participatory process: Implications for validation of standards-based accountability programs. *Educational Measurement: Issues and Practice, 21* (1), 16–22.

Sackett, D. L., Rosenberg, W. M. C., Muir Gray, J. A., Haynes, R. B., & Richardson, W. S. (1996). Evidence based medicine: What it is and what it isn't. *British Journal of Medicine,* 312: 71. Retrieved from http://www.bmj.com/content/312/7023/71.full?eaf

Texas Education Agency (2010). *House Bill 3 Transition Plan.* Retrieved from http://www.tea.state.tx.us/student.assessment/hb3plan/