



**A Standard-Setting Plan
for the State Board of Education**

January 2002

**Presented by the Texas Education Agency
Student Assessment Division
Austin, Texas**

TABLE OF CONTENTS

INTRODUCTION3

HISTORICAL PERSPECTIVE ON ASSESSMENT3

STANDARD SETTING ON TAAS4

REVIEW OF TAKS DEVELOPMENT ACTIVITIES TO DATE5

THE PURPOSE OF THE STANDARD-SETTING PLAN10

RECOMMENDATIONS OF THE TAC FOR THE PROPOSED STANDARD-SETTING PLAN11

PROPOSED BOARD ACTIVITIES AND TIME LINE15

APPENDIX A. MEMBERS OF THE NATIONAL TECHNICAL ADVISORY COMMITTEE17

APPENDIX B. THE NATIONAL TAC OVERVIEW OF STANDARD-SETTING METHODOLOGY21

Introduction

As mandated by the 76th Texas Legislature in 1999, the Texas Education Agency (TEA) is developing a new assessment program, the Texas Assessment of Knowledge and Skills (TAKS), to be administered beginning in the 2002-2003 school year. The Texas Education Code (TEC) charges the State Board of Education with establishing the passing standards (performance standards) on the new TAKS. Specifically, TEC, Section 39.024(a), mandates that “the State Board of Education shall determine the level of performance considered satisfactory on the assessment instruments.” As part of the development process for this new assessment program, TEA has asked its National Technical Advisory Committee (TAC) of educational testing experts to provide the board with information about the standard-setting process so that board members can develop the expertise necessary to perform this function. The standard-setting plan proposed here by the TAC for board approval is intended to provide a framework for a series of activities in 2002 to enable the board to be as fully informed as possible when setting the passing standards.

Historical Perspective on Assessment

The following section provides a brief history of testing in Texas with an emphasis on the standard-setting process (Cruse & Twing, 2000).¹

Texas Assessment of Basic Skills (TABS)

In 1979 the Texas Legislature passed a bill amending the TEC to require the TEA to adopt and administer a series of criterion-referenced assessments designed to assess “minimum basic skills competencies” in mathematics, reading and writing for students in grades three, five and nine. Because there was no mandated statewide curriculum at that time, the learning objectives for TABS were developed by committees of Texas educators and represented a very limited number of skills students were expected to learn in Texas public schools (Westinghouse Information Services, 1982, pp. 1-5).

Although TABS was not a diploma-denial test, ninth grade students failing to pass the TABS test were required to retake the exam each year thereafter while in school. This legislative requirement increased pressure on schools to provide remedial support for students falling below minimum expectations on TABS. In addition, the publication of campus and district results regarding student performance on TABS represented the beginning of “high stakes” accountability for large-scale assessment in Texas.

¹ For more information about the entire testing program, the reader should refer to a special edition of *Applied Measurement in Education*, 13(4).

Texas Educational Assessment of Minimum Skills (TEAMS)

In 1984 the Legislature changed the wording of the TEC, requiring the state assessment program to measure “minimum skills” rather than “minimum basic skills competencies.” This legislative mandate increased the rigor of the state assessment and added individual student sanctions for performance at the exit level. Beginning in the 1985-86 school year, the Texas Educational Assessment of Minimum Skills (TEAMS) replaced TABS as the new state-mandated criterion-referenced test in the subjects of reading, mathematics and writing. TEAMS was administered to students in Grades 1, 3, 5, 7, 9, and 11, with the 11th grade test being an “exit level” assessment. The board established passing standards for TEAMS in 1985.

Texas Assessment of Academic Skills (TAAS)

In fall 1990 changes in state law required the implementation of a new criterion-referenced program, the Texas Assessment of Academic Skills (TAAS). The implementation of TAAS shifted the focus of state assessment in Texas from minimum skills to academic skills. The TAAS tests represented a more comprehensive assessment of the state-mandated curriculum, the Essential Elements. Moreover, TAAS assessed higher-order thinking skills and problem-solving ability in reading, mathematics and writing.

TAAS is one component of a statewide-integrated school accountability system that includes the rating of campuses and districts. The Texas accountability system used to rate Texas schools includes TAAS performance results for all students and for African-American, Hispanic, White and Economically Disadvantaged subpopulations. The inclusion of TAAS in the accountability system, the public release of performance results, and the exit level requirement for graduation have made TAAS the most “high stakes” assessment in Texas history.

Standard Setting on TAAS

The *Texas Student Assessment Program Technical Digest for the Academic Year 1999-2000, Appendix 9*, provides information regarding the adoption of performance standards for TAAS. These standards were established for the school year 1990-1991 at the July 1990 board meeting. The following paragraphs provide a synopsis of this standard-setting activity.

In order to fully implement the TAAS program, the board was required to establish levels of satisfactory performance on TAAS. Districts were required by law to provide remediation to every student who failed to demonstrate satisfactory performance on a reading, mathematics, or writing TAAS test.

The board considered several important issues in the process of setting the standards on TAAS. First, TAAS assessed a broader range of the Essential Elements than TEAMS did and required students to use more higher-order thinking and problem solving as they moved from grade to grade. Second, in comparison to TEAMS, the TAAS test items were more difficult. Third, the TAAS program served multiple purposes: it was an assessment that provided scores and consequences at the student level, the school level, and the district level.

Originally the board set minimum expectations at a cut-score equivalent to approximately 70% of the multiple-choice items correct beginning with the 1991–1992 school year. The 1990–1991 school year served as a transition from the TEAMS program. The board set the interim minimum expectations standard at 65% of the multiple-choice items correct for Grades 3, 3-Spanish, and 5; and 60% of the items correct for Grades 7, 9, and 11 exit level. A student also had to score at least a 2 on the written composition in order to meet minimum expectations on the writing test.

In 1993-1994 TAAS transitioned from a fall to a spring testing program, and the assessment of reading and mathematics was expanded to include Grades 3–8 as well as exit level beginning at Grade 10. Assessing reading and mathematics in consecutive grades allowed student achievement to be compared across grades so that annual student learning progress could be measured. In January 1994 the board voted to align the passing standards at Grades 3–8 with the standard established at the exit level. This new standard, the Texas Learning Index (TLI), allowed comparisons of achievement across grades while maintaining the same passing standard for exit level students. The TLI helped districts to determine whether each student was making the yearly progress necessary to meet minimum expectations on the exit level reading and mathematics tests in 10th grade.

Review of TAKS Development Activities to Date

To provide a context for understanding the standard-setting plan, this section provides a brief historical overview of relevant TAKS development activities. These new assessments are required to be given in the 2002-2003 school year.

Regular updates on TAKS planning

As an integral part of the development and implementation of TAKS, TEA has provided regular updates on planning activities at every board meeting since this assessment legislation was passed by the Texas legislature in 1999, beginning with a special work session in November 1999. At this first meeting, Commissioner of Education, Jim Nelson, opened the session by providing the Committee of the Whole with an overview of the statewide assessment program. Educational testing experts and educators followed this presentation with more in-depth discussion. Following this work session, the agency continued these updates with regular progress reports on TAKS planning and development before the Committee of the Whole, beginning in January 2000.

In September 2001, the board adopted new rules for the assessment program, as mandated by the TEC, Chapter 39, Subchapter B. The new rules became effective November 15, 2001, and are next subject to sunset review beginning in September 2004.

The intent of the new rules, 19 TEC Chapter 101, Assessment, is not only to update the rules to reflect the recent changes in the program but also to more effectively define, reinforce, and communicate state law and rules governing the assessment program. Therefore, the current rules have been changed in the following three ways. First, the new rules include revisions to comply with changes in statute. These pertain to the implementation of the TAKS as well as the reading proficiency tests in English (RPTE), with baseline administration in spring 2000, and the state-developed alternative assessment (SDAA), with baseline administration in spring 2001. Second, the rules have been reorganized and revised to clarify the policy and standards, roles and responsibilities, and requirements of the statewide assessment program. Finally, the former rules have been revised to improve their clarity and readability for all stakeholders in public education so that the rules may more effectively promote public understanding of the assessment program and full compliance with program requirements.

Pre-Planning Advisory Group Meeting

To begin the development of a standard-setting plan for the TAKS, TEA held a “pre-planning” meeting with a group of national measurement experts during the annual conference of the Council of Chief State School Officers (CCSSO), held in Houston during the summer of 2001. This group discussed the need to understand and communicate the process used to gather evidence regarding “opportunity to learn” the knowledge and skills assessed on TAKS. Attendees also emphasized the importance of training in standard setting, not only for TEA but also for the board and others who might be involved in the standard-setting process. The pre-planning group recommended establishing a national Technical Advisory Committee (TAC) of between 10 and 20 people that would help establish the “blueprint of procedures” for the board to consider when setting standards. During this pre-planning meeting, group members discussed potential candidates to serve on the national TAC. The group felt that the national TAC should be as diverse as possible with regard to educational backgrounds and interests while still being knowledgeable of the task at hand. The group made suggestions about what materials and documentation the national TAC should review before an actual meeting was held to discuss standard-setting procedures.

Since this pre-planning meeting, TEA has been involved in the following series of discussions and work sessions with the board, including the Committee on Instruction and the Committee of the Whole.

September 6, 2001 SBOE Meeting

During this meeting before the board's Committee on Instruction, TEA presented an overview of the student assessment program. The purpose of this overview was to provide a context from which board members could effectively set passing standards for the new assessment program, the Texas Assessment of Knowledge and Skills (TAKS), in November 2002. Mr. Keith Cruse, Managing Director of the Student Assessment Division, began by outlining the history of the state assessment program and how it has evolved from the Texas Assessment of Basic Skills (TABs), administered from 1980-1985, and the Texas Educational Assessment of Minimum Skills (TEAMS), administered from 1985-1990, to the current and more rigorous Texas Assessment of Academic Skills (TAAS). The TAKS program represents a challenging but attainable next step in increasing the rigor of the Texas assessment program. Following this overview, Mr. Cruse proposed a preliminary schedule for the TAKS standard-setting activities from September 2001 to November 2002. He emphasized that this was a tentative schedule of activities that could change in response to board needs and direction. Although Mr. Cruse notified the board that the national TAC on standard setting would meet September 20, 2001, in Austin, this meeting was later postponed. Mrs. Miller, chairperson of the Committee on Instruction, agreed to this schedule and the panel and advised other board members to attend the October 31 work session with a panel representing the National TAC.

October 31, 2001 Special Work Session of the Committee on Instruction

On October 31, 2001, the Committee on Instruction held a special work session, in which a subgroup of national measurement experts from the TAC presented information critical to an understanding of the standard-setting process. The speakers at this October 31 special work session were Dr. Joanne M. Lenke, Harcourt Educational Measurement; Mr. Michael Beck, Beck Evaluation and Testing Associates; and Dr. Susan E. Phillips, Consultant. Following is a summary of each speaker's comments. (The official TEA minutes of this special work session provide detailed information.)

Dr. Lenke began her presentation by describing the distinction between content standards and performance standards. According to Dr. Lenke, content standards are synonymous with curriculum standards, which in Texas are the state-mandated TEKS. Performance standards, on the other hand, represent the performance students must demonstrate on the assessment to show that they have achieved a score (called a cut score) indicative of a pre-established level of curriculum mastery. Currently the performance standard on TAAS to pass reading and mathematics is a TLI score of 70. Dr. Lenke pointed out that some assessments are based on multiple cut scores. For example, the National Assessment of Educational Progress (NAEP) has three cut scores, or four performance levels: *Below Basic*, *Basic*, *Proficient*, and *Advanced*. Dr. Lenke added that regardless of how many levels are established, the key is to be able to adequately describe what it is that students should know and be able to do to attain each level.

Mr. Beck began by explaining that there is no one “correct” method used to set standards on an assessment. In fact, the procedures used to guide the process of standard setting vary from state to state. Based on his involvement in standard setting with numerous states, he pointed out that certain elements are common regardless of the standard-setting procedure used.

- ❑ Broad-based advisory panels are used to provide advice to the standard-setting authority. These are groups typically comprised of 20-25 people who are considered stakeholders: teachers, administrators, community and business leaders, parents, and others.
- ❑ Typically each of these advisory panels meets to discuss, in a systematic way, what the passing standard(s) should be. The purpose of such panels is to provide a forum for the consideration of different points of view during the process of drafting recommendations regarding the passing standard(s).
- ❑ There are five primary factors that the entity with the authority to establish the passing standard(s)—e.g., the board—may consider: (1) the recommendations of the various advisory panels; (2) the impact, or consequences, of various passing standards on all students as well as on different student groups (e.g., economically disadvantaged students versus non-economically disadvantaged students); (3) the consistency of the passing standard(s) both across grades and across subjects; (4) the use of various statistical measures (e.g., the standard error of measurement) in adjusting and aligning the various passing standards; and (5) the perceptions and goals of the standard-setting authority (i.e., the board).

Mr. Beck also stressed that the standard-setting process is almost always guided by an overall advisory group.

Dr. Phillips summarized the G. I. Forum litigation and outlined what information regarding standard setting could be applied to the TAKS program.² Dr. Phillips emphasized that, unlike many other aspects of testing, everything “counts” when it comes to standard setting. When a court considers whether an assessment program is equitable, all facts and circumstances regarding the standard-setting process fall under scrutiny. Dr. Phillips reiterated that, as far as the court was concerned, there is no one best way to establish passing standards. The court did not require the use of a specific standard-setting methodology; however, the court carefully reviewed what information the board considered in the process of setting passing standards. Dr. Phillips stated that having complete documentation of this information in the minutes of the board meeting(s) during which passing standards were set provided very strong evidence that the board had considered all relevant facts and circumstances.

² *GI Forum et al. v. TEA et al.*, F.Supp., 1 (W.D. Tex. 2000). Additional information can be found in the Special Issue of *Applied Educational Measurement*, 13 (4), 2000: Defending a High School Graduation Test: *GI Forum v. Texas Education Agency*, S. E. Phillips, Guest Editor.

Following the presentations by these TAC members, the board then considered components to be included in the standard-setting plan. Information collected at this October 31 work session was presented to the board at their November meeting for further input and review and then to the TAC as a basis for beginning the development of the standard-setting plan. This plan will then be presented to the board for approval in January 2002. The plan, based on input received from board members during this work session, will include, but not be limited to, the following components.

1. The plan will include a provision for the board to review the test forms for which they will be establishing passing standard(s).
2. The plan will include a provision requiring detailed impact data specifying the consequences, or impact, that various passing standards would have on Texas students. These impact data will be provided for all students as well as for each disaggregated student group as sample sizes permit.
3. The plan will include a provision requiring information regarding students' "opportunity to learn" the TEKS measured by the assessment. This information will include, but will not be limited to, results from the educator surveys collected statewide.
4. The plan will include a provision requiring recommendations from broad-based advisory panels, as described by Mr. Beck.
5. The plan will incorporate the advice of the TAC national measurement experts regarding best practice in standard setting.

November 7, 2001 Work Session

At the November meeting of the board, selected members of the TAC were asked to summarize, for the Committee of the Whole, information provided to the Committee on Instruction at the October 31 Work Session. The following paragraph outlines these presentations. (A more complete summary of these presentations can be found in the official TEA minutes.)

Dr. Lenke provided a summary of her October 31 presentation. Several members of the board who were not present for her first presentation asked questions regarding the similarities between the standard-setting processes followed by NAEP and those being planned for Texas. Dr. Jon Twing provided an overview of Mr. Beck's October 31 presentation. Some members of the board wanted to know exactly what their role was in setting standards. TEA responded that ultimately a raw score cut would be required for (each) level(s) of the passing standard(s) the board would like to set. Test-form equating would then be used to translate these cuts to future test forms. Dr. Phillips provided a more in-depth review of the GI Forum litigation and responded to numerous questions from members of the board. The meeting concluded with a presentation of the recommendations received from board members regarding components of the standard-setting plan. The board accepted these recommendations without changes or additions.

November 12, 2001 National Technical Advisory Committee Meeting

A meeting of the full body of the National TAC discussed the issues surrounding the establishment of passing standards for TAKS and advice desired from the board, as reflected in the recommendations from the November 7th Work Session. Panel members were selected based on their expertise, their diverse backgrounds and their interests. Members who did not have prior commitments participated in the meeting, which was held in Austin.³

This committee was briefed by TEA, by members who attended the board work sessions, and by the supporting contractors. The committee's primary goal was to devise a plan for the board to use that would help it during the standard-setting process.

December 4, 2001 Statewide District Advisory Panel Meeting

A meeting of district testing directors met with TEA to provide advice on various aspects of the assessment program. The advisory panel reviewed and discussed a draft version of the standard-setting plan. Suggestions from the advisory panel members have been incorporated into the standard-setting plan, as appropriate.

The Purpose of the Standard-Setting Plan

The goal of the standard-setting plan is to provide the board with the necessary information to establish performance standards that meet the needs of the State of Texas. While it is necessarily true that these standards will be based on the board's judgment, it is important that this judgment be informed. Therefore, the procedures, recommendations, and suggestions included in this plan are offered to ensure that these standards are established using nationally recognized processes and procedures. It is only through a sound methodology that the passing standards will be educationally useful as well as legally and psychometrically defensible. This defensibility will be manifested in the following ways.

1. An acceptable and tested standard-setting methodology will be used.
2. Stakeholders, including teachers, administrators, parents, and business and community leaders, will be involved.
3. The assessment on which the standards are established will be explicitly linked to curriculum.
4. Opportunity-to-learn data will be collected and used in the standard-setting process.
5. The standard-setting process will include all the facts and circumstances surrounding the assessment program.
6. The standard-setting process will be well documented and open to public scrutiny.
7. Impact of the performance standards will be carefully considered during the standard-setting process.

³ Please see Appendix A for a complete list of the National TAC and a brief description of each member's background.

Ultimately the board has the authority and responsibility to establish performance standards on the new TAKS assessments at Grades 3–11 exit level. (However, the Higher Education Coordinating Board is responsible for setting an additional standard at the exit level, which will indicate “readiness to enroll in an institution of higher education.”) It is the goal of TEA and the National TAC to help board members perform their mandated function in the best way possible.

Recommendations of the TAC for the Proposed Standard-Setting Plan

This section provides a set of recommendations that will serve as the basis for a plan that the board can use to establish performance standards for the new TAKS assessments. Following a summary list of the recommendations, each recommendation is then presented with a short discussion of critical issues and considerations. This plan is based on input from the board and advice from the national TAC. The intent of this plan is to provide milestones and a schedule outlining activities recommended by the TAC between now and November 2002, when the board will establish performance standards, in advance of the first TAKS administration in spring 2003.

The National TAC recommends the following.

1. The board should be advised by statewide standard-setting panels comprised of but not limited to Texas educators, parents, and business and community leaders to help the board make informed decisions during the standard-setting process.
2. The board should consider multiple performance standards to differentiate levels of student performance.
3. The board should consider developing a system for phasing in performance standards for certain subject areas, grades, or purposes.
4. The board should establish distinct phases for the standard-setting process:
 - a. adopt the standard-setting plan
 - b. authorize TEA to convene statewide standard-setting panels
 - c. receive information on standard-setting methodologies
 - d. review secure test forms
 - e. receive recommendations from statewide standard-setting panels
 - f. consider impact data and other information that would inform the board’s decisions
 - g. set performance standards

Recommendation 1:

The board should be advised by statewide standard-setting panels comprised of but not limited to Texas educators, parents, and business and community leaders to help the board make informed decisions during the standard-setting process.

The National TAC recommends that the statewide standard-setting panels be comprised of diverse groups of stakeholders. The committee believes that involving a variety of stakeholders assures that different perspectives will be represented during the standard-setting process. In addition, the inclusion of diverse groups in the decision-making process will make it more likely that the public will accept the performance standards the board adopts.

The statewide standard-setting panels should represent multiple grades within specific content areas. This configuration will address one of the challenges noted by the national TAC: the need to keep the standards at adjacent grades within a content area consistent with one another. Without such consistency, the assessment program will lack the stability necessary to be a valid and reliable measure of student performance from year to year.

The standard-setting panels may address the following issues:

- what constitutes “satisfactory” performance
- whether multiple performance levels should be established and, if so, how these performance levels should be described
- how standards might be phased in over time

Recommendation 2:

The board should consider multiple performance standards to differentiate levels of student performance.

During the October and November work sessions, members of the board expressed an interest in establishing more than a single cut score, or “pass/fail” standard. Many state assessments currently have multiple levels of proficiency. In addition, the National Assessment of Educational Progress (NAEP) has three cut scores that yield four levels of achievement. Furthermore, the federal reauthorization of Title I (Public Law 103-382) of the Elementary and Secondary Education Act (ESEA) requires states to establish at least three levels of student performance on their state assessments for Title I purposes. While this Title I requirement is not the focus of the current activities, the board should consider it when setting standards on TAKS.

The National TAC noted that the number of students classified in each proficiency category is an important consideration in establishing multiple passing standards. For example, if too few students are likely to fall into a particular category, then that category may serve no useful purpose. On the other hand, the board should recognize that performance is likely to improve over time, resulting in student movement from one category to another. Consequently, while there may be few students in a high proficiency category at the onset of the new assessment, more and more students are likely to achieve that level in the future, thereby improving the utility of this category over time.

As discussed previously, Section 39.024(a) of the TEC requires that the board must establish “the level of performance considered to be satisfactory.” This means that one level is required (i.e., one cut score with two categories—pass/fail); however, the law does not prohibit the establishment of other performance levels. At the exit level, the law requires that two cut scores and three categories be established: the pass/fail standard (representing the first two categories) set by the board and the “higher education component” (representing the third category) set by the Higher Education Coordinating Board.

The number and difficulty of the questions on each test will have an impact on the number of performance levels the test will ultimately be able to support. For example, a relatively short test may not be sensitive enough to provide data to accurately classify students into multiple categories. In addition, when making the decision regarding the establishment of multiple performance levels, the board must consider two things: the standard error of measurement at the cut scores and the degree to which the performance levels yield valid and reliable results.

The descriptions of student behavior and the label attached to each performance level are important considerations in establishing multiple performance levels. For example, if three levels are desired, there are likely to be different perceptions if the performance levels are labeled “Below Standard,” “Meets Standard,” and “Exceeds Standard” versus “Needs Intervention,” “Meets Expectations,” “Distinguished Achievement.” The national TAC pointed out that some states have even chosen to avoid labels altogether and have instead numbered their performance levels—Level I, Level II, Level III, and so on—to avoid any value statements and to allow more detailed descriptive statements to define what each performance level means. There are an infinite number of such labels, all of which have benefits and weaknesses and communicate slightly different messages about student learning and progress.

The National TAC offers the following advice to the board with regard to multiple performance descriptions.

1. The descriptions should accurately reflect the labels attached to the performance levels.
2. The descriptions should be short, concrete statements about what it means for a student to be at a particular proficiency level.
3. The descriptions should provide schools with information that helps them improve instruction so that students are able to move to higher levels of achievement.

Recommendation 3:

The board should consider developing a system for phasing in performance standards for certain subject areas, grades, or purposes.

The primary purpose of performance standards is to communicate the expected level of achievement to students, schools, parents, and the public. As such, these standards are really goals that define the level of performance necessary for students to make sufficient academic progress from year to year. For example, it would be of little use to establish performance standards that would result in 100 percent classification of students in the highest (or lowest) category. If performance standards are to have the desired effect—i.e., strengthen instructional programs as well as improve student achievement—these standards must differentiate among students with regard to their individual performance on the assessment. To lessen the impact of performance standards on individual students or student subpopulations and give schools the time needed to strengthen instructional programs, the board may consider adopting a system that establishes graduated performance standards. In such a system, the board may establish standards that are phased in over one or more years. Phasing in performance standards may be particularly appropriate for subject areas or grades not previously assessed at the state level.

Recommendation 4:

The board should establish distinct phases for the standard-setting process:

- a. adopt the standard-setting plan
- b. authorize TEA to convene statewide standard-setting panels
- c. receive information on standard-setting methodologies
- d. review secure test forms
- e. receive recommendations from statewide standard-setting panels
- f. consider impact data and other information that would inform the board's decisions
- g. set performance standards

Proposed Board Activities and Time Line

The standard-setting plan outlines the activities required for the board to carry out their responsibility to establish performance standards on the new TAKS assessments. These activities will be accomplished through a series of board work sessions as well as through a number of standard-setting advisory panel meetings prior to the November 2002 board meeting. The following time line summarizes the key phases of the standard-setting process designed to provide the board with the information it needs to establish fair yet challenging standards for Texas students.

September 2001 SBOE Meeting

TEA presented an overview of the student assessment program to the Committee on Instruction and proposed a preliminary schedule for the TAKS standard-setting activities from September 2001 to November 2002.

October 2001 Work Session

Nationally recognized measurement experts provided information on the standard-setting process to the Committee on Instruction and other board members.

November 2001 SBOE Meeting

Members of the national TAC provided an overview of the topics presented to the Committee on Instruction during the October work session. Board members offered recommendations they wanted added to the standard-setting plan.

January 2002 SBOE Meeting

Members of the national TAC present the Plan for Standard Setting to the board for approval.

March 2002 SBOE Meeting

The board receives information on how the TAKS items are aligned with the TEKS.

May 2002 SBOE Meeting

Members of the national TAC provide information to the board on the “item-mapping” procedure for standard setting. See Appendix B for more information.

July 2002 SBOE Meeting

Members of the national TAC provide information to the board on the data needed to make standard-setting decisions.

September 2002 SBOE Meeting

The board reviews actual TAKS test forms.

October 2002 Work Session

The board receives recommendations from the standard-setting advisory panels. In addition, board members will have an opportunity to address any unresolved issues with members of the national TAC prior to the completion of the standard-setting process in November 2002.

November 2002 SBOE Meeting

The board considers the results of the impact data from the fall 2002 study, the spring 2002 field test, recommendations from the standard-setting advisory panels, and the field-test data that link TAKS to TAAS. The board officially determines the performance standard(s) as recorded by vote.

Appendix A. Members of the National Technical Advisory Committee

Michael Beck

As founder and president of Beck Evaluation and Testing Associates (BETA, Inc.), Michael Beck has planned and directed the standard-setting process for 14 state-level assessment programs in 10 states over the past seven years and has served as a consultant in the process for five other states. Overall, BETA, Inc., has developed assessments for 18 programs in 11 states and provided consultation on test-development and test-interpretation issues to 22 state departments of education. In the past year, BETA, Inc., has developed more than 29,000 multiple-choice and open-ended items for use by state and textbook-publisher clients. Before establishing BETA, Inc. in 1983, Mr. Beck was employed by The Psychological Corporation, where he was involved in all phases of test development.

Gregory Cizek

Gregory J. Cizek is Associate Professor of Educational Measurement and Evaluation at the University of North Carolina-Chapel Hill, where he teaches courses in applied psychometrics, statistics, and research methods. Previously Dr. Cizek provided management and psychometric expertise for American College Testing (ACT) for licensing and certification programs, with a primary emphasis in the health-profession fields. Working with ACT, he also helped develop a standard-setting procedure for use with the National Assessment of Educational Progress (NAEP). For the past decade, he has provided standard-setting consultation related to large-scale state testing programs in the elementary and secondary grades. He is the editor of the *Handbook on Educational Policy and Setting Performance Standards: Concepts, Methods, and Perspectives* and has written a number of articles for psychometric journals.

Barbara Dodd

Dr. Barbara Dodd is a Professor in the Department of Educational Psychology at the University of Texas at Austin, where she received her Ph.D. in 1984. Her research interests include the application of classical mental test theory and item-response theory to attitude scaling, academic placement, test equating, and adaptive testing. She currently serves on the Technical Advisory Committees for several state testing programs.

Richard Duran

Dr. Richard Duran received his Ph.D. in Quantitative Psychology from the University of California-Berkeley in 1977. Prior to his appointment as a faculty member in the Graduate School of Education at the University of California-Santa Barbara, he served for seven years as a research scientist as well as the Coordinator of Research for the Test of English as a Foreign Language for the Educational Testing Service (ETS) in Princeton, N.J. In addition to his work at ETS, Dr. Duran has conducted numerous investigations of the test performance and academic achievement of students from non-English backgrounds. He has published two books and numerous articles on the testing of linguistically diverse students. He has served as a member of the National Research Council Board on Testing and Assessment and the U.S. Defense Department Advisory Committee on Personnel Testing. Currently he serves on external Technical Advisory Committees for the state assessment offices for Oregon, Washington, and New York. He also serves on the National Assessment of Educational Progress (NAEP) Validity Studies Panel.

David Francis

Dr. David J. Francis is a professor of Quantitative Methods in the Department of Psychology and the Director of the Texas Institute for Measurement, Evaluation, and Statistics at the University of Houston. He is a fellow of Division 5 (Measurement, Evaluation, and Statistics) of the American Psychology Association and a current member of the National Institute of Child Health and Human Development (NICHD) Mental Retardation Research Subcommittee. Dr. Francis has collaborated in research on reading and reading disabilities, attention problems, developmental consequences of brain injuries and birth defects, and adolescent alcohol abuse. In addition, Dr. Francis currently serves as consulting editor to several psychological journals and is founding partner of FSD Data Services, a contract research services firm based in Houston.

Joanne Lenke

Dr. Joanne Lenke has been involved in the test publishing industry for more than 30 years, both as an independent consultant and as an employee of Harcourt Educational Measurement (HEM). As part of her duties, she has participated in and supervised the research and development of a variety of both norm- and criterion-referenced tests. In addition, she designed the first standard-setting study conducted for a norm-referenced test and assisted in training and facilitating the standard-setting panels for that test. For the past two years, Dr. Lenke has served as a consultant specializing in testing issues. Currently she serves as Vice-President of Custom Test Programs for HEM.

William A. Mehrens

A professor emeritus of the Department of Education at Michigan State University in East Lansing, MI., Dr. William A. Mehrens is considered one of the foremost experts in the field of psychometric measurement theory and education. He has authored numerous books used widely as college texts in educational measurement and has won a variety of honors for his contributions in the field. His interests include legal issues in high-stakes testing, teaching to the test, and performance assessment. He has held office in several professional organizations, including as president of the National Council on Measurement in Education (NCME), as president of the Association for Measurement and Evaluation in Guidance, and as vice president of Division D of the American Educational Research Association (AERA). Dr. Mehrens served as a professor at Michigan State University from 1970 until his retirement last year.

Susan Phillips

A former professor of Educational Measurement at Michigan State University and an attorney, Dr. S. E. Phillips is one of the nation's foremost authorities on assessment law, particularly with regard to high-stakes statewide assessments and standardized tests. Recently Dr. Phillips served as an expert witness and consultant in the Texas GI Forum lawsuit, in which the state of Texas successfully defended the exit level TAAS test required for high school graduation. Dr. Phillips also serves on Technical Advisory Committees or as a consultant on legal or psychometric for numerous other statewide testing programs in which standard setting is a primary issue. Dr. Phillips has served as a consultant to TEA for statewide testing programs since the early 1980s.

Michael Rodriguez

As a professor of Educational Psychology at the University of Minnesota, Dr. Michael C. Rodriguez has participated in large-scale test design for both regular and special education settings in several states, with a special interest in alternative assessments for students with moderate to severe disabilities. Recently he has worked with the Educational Testing Service (ETS) to examine possible impact related to cut scores for a principal's licensure program. Currently he is investigating variability in measurement error of state test results across schools and is involved in a research project examining Latino youth development.

Joseph Ryan

Director of the Research Consulting Center at Arizona State University West in Phoenix and a professor at the university, Dr. Joseph Ryan has worked in the area of applied psychometrics for more than 20 years. He has extensive experience with the standard-setting process and procedures, including the Nedelsky, Angoff, and various other item-mapping methodologies. Recently he contributed two chapters to the book, *Issues, Research, and Methodologies for Large Scale Assessment Programs*, G. Tindel and T. Haladyna (Eds.), to be published by Lawrence Earlbaum Co. Dr. Ryan has also served as an advisor on standard setting in a dozen states and has worked as a consultant for most of the major testing companies.

E. Roger Trent

Dr. E. Roger Trent is Executive Director Emeritus for the Ohio Department of Education. Prior to this, Dr. Trent served as a director for 14 years. In these roles Dr. Trent was responsible for the overall operation of statewide testing programs in Ohio. His duties included recommending to the state board of education performance standards for proficiency tests in five subject areas at four grade levels; working with a national technical advisory committee to define a standard-setting methodology; convening committees to recommend standards for each test; convening review committees to determine whether the proper methodology was followed and whether standards appeared to be reasonable; and recommending specific cut scores for each test. For more than a decade, Dr. Trent has participated as a member of Technical Advisory Committees in a number of states.

Audrey Qualls

Dr. Audrey L. Qualls is an associate professor of educational measurement and statistics at the University of Iowa and a co-author of the Iowa Test of Basic Skills. Her research and expertise include the development of large-scale assessment tools, score reliability, and the interpretation and use of information yielded from standardized tests. She teaches both undergraduate and graduate courses in applied statistics and educational measurement. Dr. Qualls has a Ph.D. in educational measurement and statistics from the University of Iowa.

Appendix B. The National TAC Overview of Standard-Setting Methodology

There are a variety of standard-setting methods, all of which require the judgments of educational experts and possibly other stakeholders. These experts are frequently referred to as judges, participants, or panelists (the term panelist will be used here). Acceptable methods for standard setting could be test-centered or student-centered (Jaeger, 1989). Test-centered methods focus panelists' attention on the items in the test. Panelists make decisions about how important and/or difficult the test content is and make judgments based on that importance. Student-centered methods focus panelists' attention on the actual performance of examinees or groups of examinees. Cut scores are set based on student exemplars of different levels of competency. In addition, standards can be set using either a compensatory or conjunctive model (Hambleton & Plake, 1997). Compensatory models allow examinees who perform less well on some content to "make up for it" by performing better on other important content. Conjunctive models require that students perform at specified levels within each area of content.

Many standard-setting methods are better suited to specific conditions and certain item types. For example, the popular Modified Angoff method appears to work best with selected-response (SR) items (Cizek, 2001; Hambleton & Plake, 1997), while the "judgmental policy-capturing method" was designed specifically for complex performance assessments (Jaeger, 1995). Empirical research has repeatedly shown that different methods do not produce identical results, and it is important to consider that many measurement experts no longer believe that "true" cut scores exist (Zeiky, 2001). Therefore, it is crucial that the method chosen meet the needs of the testing program.

The Standards for Educational and Psychological Testing (AERA, APA, NCME, 1999) detail issues that should be addressed in all educational testing situations. While not specifically addressing standard setting, several standards are relevant. Table 1 presents the applicable standards (paraphrased in Zeiky, 2001).

Table 1. Joint Technical Standards

Standard	Topic
2.14	The standard errors of measurement should be reported near the cut score.
2.15	The percent of test takers should be classified the same way on repeated measures.
4.4	The score interpretations based on cut scores should be described and justified in the same way.
4.19	The rationale and procedures for setting cut scores should be documented.
4.20	When feasible, cut scores “should be established on the basis of sound empirical data concerning the relation of test performance to relevant criteria” (<i>Standards</i> , p. 60).
4.21	If judgments of test items are used, “the judgmental process should be designed so that judges can bring their knowledge and experience to bear in a reasonable way” (p. 60).
6.5	Test documentation should include material on the test’s reliability and validity, the equating procedures used with the test, performance levels and cut scores associated with the test, and standard errors of measurement.
6.12	Users should be informed of cut scores used in computer-generated interpretations of scores.
13.6 ⁴	For promotion or graduation tests, multiple opportunities to pass or alternative measures should be available.

⁴ For TAKS, there will be multiple opportunities to pass, with mandatory remediation between each opportunity to retake the test, before a student will be denied a high school diploma.

Methods

A multitude of different procedures or methodologies can be used to establish performance standards. Two of the most often used are the Modified Angoff Procedure and an item-mapping procedure.

The Modified Angoff standard-setting procedure is both test-centered and compensatory. In the Modified Angoff, panelists first operationally define the prescribed proficiency levels by referring to behavioral expectations of the standards (i.e., the descriptors) and their own experience with student work. Once panelists have agreed on these definitions, they are then asked to consider the performance of 100 students who just barely meet expectations at each proficiency level. These students comprise “borderline” groups. Each panelist then works independently to assign what he/she believes to be the number of students in each borderline group that will answer each item on a test correctly.⁵ These numbers are values reflecting the proportion of borderline students correctly responding to each item. A statistical procedure is then used to derive a cut score. This procedure is repeated once or twice, with groups holding discussions between iterations, or rounds, in order to encourage consensus.

The Modified Angoff was once seen as “best practice” and has resulted in standards that are legally defensible. However, this procedure has drawbacks. First, it does not work well with constructed-response (CR) items. This is because successive score levels may require different degrees of specific knowledge or skills from students. For example, on a 3-point mathematics item, the mathematics achievement required of students to move from a score of 1 to 2 may be less than that required of students to move from a score of 2 to 3. Therefore, to be used on a large-scale assessment that includes open-ended items, the Modified Angoff procedure would require modifications. This drawback would, by necessity, have to be considered when setting standards on TAKS, since the Grade 9 reading test and the Grades 10 and 11 exit level English language arts tests include performance items. Secondly, the Modified Angoff method has been criticized as being too cognitively demanding of panelists (National Academy of Education, 1993) because they must keep in mind their definition of a borderline group (for each performance level) and then inspect each test item and estimate the percentage of borderline students answering that item correctly.

A procedure that is being used more and more widely in the industry is the item-mapping method. While this method is relatively new, it is quickly gaining acceptance as an alternative to the Modified Angoff, especially for tests containing performance-based items (Zeiky, 2001). This method is also both test-centered and compensatory. Proponents of item mapping claim it has two major advantages over the Modified Angoff: (1) item mapping is less cognitively taxing on panelists, and (2) it integrates the treatment of multiple-choice and performance-based items (Mitzel, Lewis, Patz and Green, 2001).

⁵ Panelists also consider impact data (both overall and separately by minority group), actual student performance on each item, other judges’ ratings, as well as other information.

The item-mapping standard setting procedure is similar to the Modified Angoff in terms of logistics and personnel requirements but differs in terms of the panelist activities. Rather than having panelists estimate student performance on the items, item mapping presents the items in actual order of difficulty. Rather than estimating the proportion of borderline examinees responding correctly, panelists place a “mark” between the hardest item a borderline examinee should answer correctly and the easiest item the examinee should miss. This point defines the cut score. Although the Modified Angoff also includes a social component, the social activities included in the item-mapping method are relatively more prescribed. For example, panelists are instructed to discuss the elements of each item that makes that item more difficult than the items that precede it.

In item mapping, after all items included in the standard setting have been calibrated using an Item Response Theory measurement model, an item map is created. The item map arranges items according to empirical difficulty and lists the following information: rank difficulty, item type, the place in which the item appears in the test, the answer key (for constructed-response items, the number of score categories), the difficulty location for the item, and the item-content classification. An item book, which presents all the items—one per page—in order of difficulty, is prepared. One drawback to the item-mapping procedure is the tendency for items to be judged as “disjointed” by members of the committee. For example, judges sometimes disagree on the order of difficulty of the items. One judge may believe that a particular item is too hard for a borderline student but that the item after it may be easy enough for the student to answer, while another judge may agree with the original order of items with regard to difficulty. This type of disagreement may arise from a variety of factors: measurement error, situational or specific knowledge, and inaccurate perceptions on the part of some panel members. This is one reason a broad-based, diverse group of multiple judges is used.

References

- American Educational Research Association, American Psychological Association, & National Council of Measurement in Education (1999). *Standards for Educational and Psychological Testing*.
- Phillips, S. E., (Ed.) (2000). Special Edition of *Applied Measurement in Education*, (13) 4.
- Cruse, K. L., and Twing, J. S., (2000), The history of statewide achievement testing in Texas. *Applied Measurement in Education*, (13) 4, pp. 327-332.
- Cizek, G. (2001). Conjectures on the rise and call of standard setting: An introduction to context and practice. In G. Cizek (Ed.), *Setting Performance Standards: Concepts, Methods, and Perspectives* (pp. 3-17). Mahwah, NJ: Erlbaum.
- Cizek, G., (Ed., 2001). *Setting Performance Standards: Concepts, Methods, and Perspectives*, Mahwah, NJ: Erlbaum.
- Department of Children, Families and Learning & NCS Pearson (2001). *Minnesota Comprehensive Assessments: Technical Manual*.
- Hambleton, R. & Plake, B. (1997). An anchor-based procedure for setting standards on performance assessments. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.
- Hambleton, R. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. Cizek (Ed.), *Setting Performance Standards: Concepts, Methods, and Perspectives* (pp. 89-116). Mahwah, NJ: Erlbaum.
- Huynh, H. (2000). On Bayesian rules for selecting 3PL binary items for criterion-referenced interpretations and creating booklets for bookmark standard setting. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.
- IOX Assessment Associates. (1985). *Setting passing standards for the Texas Educational Assessment of Minimum Skills*. Culver City, CA: IOX Assessment Associates.
- Jaeger, R. (1995). Setting standards for complex performances: An iterative, judgmental policy-capturing strategy. *Educational Measurement: Issues and Practice*, Winter, 16-20.
- Jaeger, R., (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational Measurement*, (3rd Ed., pp. 485-514). Washington DC: American Council on Education.
- Kane, M. (2001). So much remains the same: Conception and status of validation in setting standards. In G. Cizek (Ed.), *Setting Performance Standards: Concepts, Methods, and Perspectives* (pp. 53-88). Mahwah, NJ: Erlbaum.
- Lee, G. & Lewis, D. (2001). A generalizability theory approach toward estimating standard errors of cut scores set using the bookmark standard setting procedure. Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement*, (3rd Ed., pp. 13-103). Washington DC: American Council on Education.

- Mitzel, H., Lewis, D., Patz, R. & Green, D., (2001) The bookmark procedure: Psychological perspectives. In Cizek, G. (Ed.), *Setting Performance Standards: Concepts, Methods, and Perspectives* (pp. 249-281). Mahwah, NJ: Erlbaum.
- Raymond, M. & Reid, J., (2001). Who made thee a judge? Selecting and training participants for standard setting. In G. Cizek (Ed.), *Setting Performance Standards: Concepts, Methods, and Perspectives* (pp. 119-158). Mahwah, NJ: Erlbaum.
- Reckase, M. (2001). Innovative methods for helping standard-setting participants to perform their task: The role of feedback regarding consistency, accuracy, and impact. In G. Cizek (Ed.), *Setting Performance Standards: Concepts, Methods, and Perspectives* (pp. 159-174). Mahwah, NJ: Erlbaum.
- Texas Education Agency (TEA). (1990). *Texas Educational Assessment of Minimum Skills and Texas Assessment of Academic Skills: Annual report*. Austin, TX: National Computer Systems.
- Texas Education Agency (TEA). (1994). *Texas Student Assessment Program Technical Digest for the Academic Year 1993-1994*. Austin, TX: National Computer Systems.
- Texas Education Agency (TEA). (1996). The development of accountability systems nationwide and in Texas. Publication Number GE6-601-07. Austin, TX: The Texas Education Agency.
- Texas Education Agency (TEA). (2001). *Accountability Manual*. Publication Number GE01-602-03. Austin, TX: The Texas Education Agency.
- Westinghouse Information Services. (1982). *Texas Assessment of Basic Skills 1981-1982 Final Report: Project Report*. Iowa City, IA: Westinghouse Information Services.
- Zieky, M. (2001). So much has changed: How the setting of cut scores has evolved since the 1980s. In G. Cizek (Ed.), *Setting Performance Standards: Concepts, Methods, and Perspectives* (pp. 19-52). Mahwah, NJ: Erlbaum.