Summary Report

Setting Student Performance Standards for the Texas Assessment of Knowledge and Skills (TAKS)

Report Prepared by
BETA, Inc.

November 2002

This report summarizes the activities leading to the establishment of student performance standards for the new Texas statewide assessments, the Texas Assessment of Knowledge and Skills (TAKS). Most of these activities were conducted by Beck Evaluation & Testing Associates, Inc. (BETA) under contract with the state's prime contractor for the TAKS program, NCS Pearson. This Summary Report is divided into five sections to describe the major stages of the standard-setting process:

- Develop a research-based plan
- Choose and describe the categories used to label student test performance
- Convene panels to recommend standards for each TAKS test
- Review all panel-recommended standards
- Convene panels to recommend standards for each Spanish TAKS test
- Present all recommendations to the State Board of Education (SBOE)

Develop a Research-Based Plan

The overall TAKS standard-setting process began with a "pre-planning" meeting with a group of national measurement experts, which was convened during the Council of Chief State School Officers (CCSSO) annual conference in Houston in June of 2001. This group discussed the need to ground the process in the advice of a technical advisory group of national experts. In addition, the group recommended the development of a written plan to guide all aspects of the standard-setting process; this plan would be submitted for SBOE approval.

Subsequent to that 2001 meeting, the Texas Education Agency (TEA) assembled and convened the Technical Advisory Committee (TAC), as recommended. This committee, composed of 13 national experts in assessment, has met six times (four times in Austin and twice via telephone conference call) to react to various aspects of the

standard-setting plan and process and to advise TEA on other relevant issues. A complete list of TAC members is presented in Appendix A. Several members of that committee have made presentations to the SBOE in five separate 2001–2002 sessions on various aspects of the TAKS standard-setting process.

In early fall of 2001, TEA staff and their contractors and consultants drafted an overall Standard-Setting Plan to guide this process. This plan went through several iterations with revisions made on the advice of the SBOE, the TAC, and other TEA advisors. The final document, *A Standard-Setting Plan for the State Board of Education*, was presented to and adopted by the SBOE at their January 2002 meeting. This plan was used to direct all efforts leading to the recommended set of standards. The SBOE-approved plan is presented in Appendix B of this report.

Following SBOE approval of the overall plan for the process, TEA and its consultants drafted an *Implementation Plan*. This plan was revised several times during the spring and early summer of 2002. It details the specific activities conducted to arrive at a set of recommended performance levels to propose to the SBOE. That plan is presented in Appendix C of this summary.

Choose and Describe Labels for Student Performance

The critical step of selecting and describing the various student-performance "labels" to report student performance on the TAKS was assigned to a specially convened Standard Setting Advisory Panel. This committee was composed of 19 individuals representing important state professional, educational, or public-policy organizations. A list of these panelists and the agenda for their meeting are presented in Appendix D.

This committee's primary function was to select appropriate descriptive labels for the student-performance categories and to set forth broad, generic descriptors for each label. The panel met for a full day in Austin on May 23, 2002; their session was facilitated by representatives of BETA. TEA staff participated as observers and resource personnel; they took no active role in the discussions or in the selection of the terms or descriptors.

The panel was presented with a broad range of eighty potential labels. Panelists then spent several hours participating in activities designed to reduce the potential label candidates into smaller and smaller lists. After discussion and selection of the three preferred labels, they generated various phrases and descriptors for inclusion with the labels. The labels and descriptors chosen were intended to be equally applicable to all grade levels and content areas covered by TAKS. The labels recommended by the Advisory Panel, with the generic descriptors added by TEA based on the recommendations of the Advisory Panel, were as follows.

Commended Performance

- *High* academic achievement
- Students performed at a level that was considerably above the state passing standard
- Students demonstrated a *thorough* understanding of the knowledge and skills measured at this grade

Met the Standard

- *Satisfactory* academic achievement
- Students performed at a level that was at or somewhat above the state passing standard
- Students demonstrated a *sufficient* understanding of the knowledge and skills measured at this grade

Did Not Meet the Standard

- *Unsatisfactory* academic achievement
- Students performed at a level that was below the state passing standard
- Students demonstrated an *insufficient* understanding of the knowledge and skills measured at this grade

These labels and a description of this process were presented to the SBOE for its review at the July 2002 meeting.

When each of the standard setting panels met, a significant activity that took place prior to the panels making any recommendations was for them to extend the above generic descriptors to the specific grade level and content area on which they were to recommend standards. This activity permitted panelists to conceptualize more clearly the various labels in terms of specific grade- and content-based behaviors.

Convene Panels to Recommend Standards for the English TAKS Tests

The large bulk of the actual effort that led to the recommended standards was conducted by 21 panels—5 of which recommended standards for English versions of TAKS and six for Spanish TAKS.

A total of more than 270 panelists participated in the English standard-setting sessions. The number of representatives per panel ranged from 15 to 22. Panelists were selected and invited by TEA based on recommendations received from the legislature, the SBOE, TEA staff, and school-district personnel. The majority of the panelists on each committee were active Texas educators–either classroom teachers at or

adjacent to the grade level for which the standards were being set, or campus or district administrative staff. All panels included representatives of the community "at large;" that is, panelists who were not practicing K–12 educators. No panelists had participated in earlier TAKS item or prototype item development activities. Travel costs were reimbursed for each attending panelist. Appendix E contains a list of all panelists.

Standard-setting sessions were either two-day or three-day meetings. Panelists participating in the three-day sessions recommended standards for two tests; those participating in the two-day sessions recommended standards for a single test. Panelists clearly understood that their role was that of an advisory group—to recommend a set of standards to TEA and the SBOE. Panelists were also told that a subgroup of each panel would be invited to attend a later one-day "review" session to discuss all standards across grades within that particular content area; these review sessions were convened at the end of all grade-specific sessions.

As recommended by the standard-setting contractor and approved by the TAC and SBOE, the general methodology used for the sessions was "item mapping." This item-mapping method, initially proposed by CTB/McGraw-Hill and termed the "Bookmark Procedure" (c.f., Mitzel, Lewis, Patz, & Green, 2001; Lewis, Green, Mitzel, Baum, & Patz, 1998), was chosen for several reasons. First, it is currently the most widely used method for high-stakes educational assessment standard setting and is used in the majority of statewide testing programs for which student performance standards are determined by panels. Therefore, it is widely understood by measurement professionals. Second, it is a procedure well suited for assessments that contain both selected-response (multiple-choice) and constructed-response items. While only some of the TAKS tests contain constructed-response items, it was considered very desirable to use a single methodology for all tests.

It is important to note that the actual methodology used was item mapping; certain methodological aspects of the Bookmark Procedure, as described by its developers, were not implemented. For example, as depicted in the literature, panelists using the Bookmark Procedure break into small subgroups to discuss their interim ratings. All panel discussions for TAKS involved the complete panel; panelists were asked not to discuss their judgments other than with the panel as a whole. In this respect, the TAKS process more closely resembled a jury process than does the original Bookmark Procedure. In addition, a probability value of 67% is typically used in the Bookmark Procedure (although variations in this probability value are not unusual).

For the TAKS standard setting, panelists were trained to examine each item, which had been ordered in a review booklet from least to most difficult. Panelists progressed through the booklet until they reached the point at which they believed a threshold student who minimally Met the Standard should *more likely than not* be able to answer the item correctly. That is, panelists placed a cut point at the item at which a student who answered correctly was just *barely* indicating performance that Met the Standard. A similar process was then followed for the Commended cut. Panel discussions between rounds of ratings indicated that some panelists appeared to have

difficulty with this determination at the Commended level. In such cases these panelists seem to have placed their cut at the point they believed a threshold Commended student should answer all items up to that point correctly and items beyond that point incorrectly (c.f., Buckendahl, Smith, Impara, & Plake, 2002). For additional information and data concerning methodological procedures and session outcomes, see the report, *Setting Standards on the TAKS Tests: A Modified Item Mapping Procedure (2002).*

Agendas for both the three-day and two-day sessions are presented in Appendix F, which also shows the schedule for each session. All panels were facilitated by a contractor's staff member with prior experience moderating standard-setting sessions for high-stakes testing programs. Facilitators all followed the same agenda and used the same overhead transparency sequence and notes to lead their sessions. All sessions within a content area were led by the same facilitator to minimize any potential "facilitator effects" on the process. Appendix G contains the transparencies used for all sessions. A single staff member from TEA was present in each session but did not participate in standards-based discussions of the panels, nor did they offer opinions as to the appropriate standards; TEA staff served only as a panel resource when TAKS program-based questions arose. Several sessions were attended by representatives of the state's TAC, who also served only as observers.

As indicated in the session agendas, each panel received extensive training or orientation prior to making any recommendations. This training, more than one-half day in duration, included an overview of the general process for establishing performance standards, a discussion of the labels to be used for characterizing student performance, and an extensive discussion regarding how the generic label descriptors could be concretely translated into student performance at the particular content area and grade for which each panel was setting standards. Following the training, each panelist had the opportunity to review, or "take," the actual TAKS test and to discuss the content with fellow panelists. Finally, panelists received an orientation to the item mapping procedure and practiced the mechanics of the process using a short "practice test" composed of TAKS prototype items in the particular content area.

The actual standard setting activities involved three rounds of recommendations. The first round took place following the aforementioned training process. Prior to the second round, panelists were shown information concerning their fellow panelists' anonymous preliminary judgments, along with summary data describing these judgments. They also received item-difficulty estimates for each item on the test. Panelists then had an extensive total-group discussion of their preliminary ratings, grounded in the particular test and items on which the standards were being set. Facilitators attempted to include all panelists in the discussions and to provide an opportunity for full discussion of the rating process and judgments. Following the second round of ratings, panelists again saw the judgments of their peers and discussed their ratings; they also were shown projected "impact data" based on the Round 2 median cuts of the panel. These impact data included the projected percent of students who would attain scores on the TAKS that either Met the Standard or were Commended. The projected impact data were generated on the basis of the spring 2002 TAKS field test.

It is important to note that the spring 2002 field test was conducted under conditions in which students were generally aware of the fact that their performance did not "count"—that is, the field-test forms were clearly identified as such and were administered shortly after the actual, operational TAAS tests were taken. The lack of student (and teacher/school) motivation to exert maximum effort in engaging the field-test questions was a significant point of discussion among panelists. Panelists considered this issue most salient at the higher grades, i.e., Grade 7 and higher.

The projected impact data were presented for the statewide overall population of Texas students; projected impact data for gender and major ethnic groups were also available. Following another set of group discussions, panelists then were given unlimited time to independently complete their third and final round of ratings. Once the final rating for a test was turned in, each panelist also completed a short Evaluation Form (adapted from a form presented by Hambleton, Jaeger, Plake, & Mills, 2000) that encouraged them to express their views concerning various aspects of the session, their comfort with their recommendations, and related issues.

Appendix H presents summary data by round for each session. One page of data is presented for each grade and content area. Summarized are the percent of Texas students projected (based on the spring 2002 field test) to score in each performance category, the panel-recommended raw-score cuts (median, mean, and standard deviation) by round, and several estimates of error related to the Round 3 recommendations. The errors presented include the standard errors of the mean and median cut scores; these provide an indication of the amount of error associated with the average (mean or median) cut scores recommended by the panel. Because of the size of the panels and their general agreement as to the cut scores (as reflected by the Round 3 standard deviations), most of these standard errors were on the order of 1 raw score point or less.

Also provided on these summary pages are estimates of the standard error of measurement of the test; these give an indication of the amount by which an individual's score on the test is likely to vary due to chance considerations. The standard error of measurement is essentially an expression of the test's reliability expressed in terms of an individual student's score. The final standard error presented on these pages combines the standard error of the test and that of the median cut score. (The median cut scores recommended by the panel were used as each panel's recommendation. Medians were selected because medians, as a measure of central tendency, are less impacted by extreme judgments than are means.) This combined standard error term, being a combination of both the error of the judges and that of the individual test-taker, is considered to be the standard error term that best captures the overall error in the standard setting and testing process.

A scan of the English TAKS test pages in Appendix H leads to several summary conclusions across content areas and grades:

- There was a fair amount of variation across grades and content areas in the projected statewide impact of the panel's recommended cuts. For example, the (Round 3) projected percent of Texas students scoring Below the Standard ranges from a low of 10% (Grade 8 Social Studies) to a high of 72% (Grade 10 Mathematics). Similarly, the projected percent of Commended students across the content areas and grades ranged from 2% to 34%. These cut scores represent a panel's subjective view of the level of performance that a threshold student should be able to demonstrate based on tests that vary in difficulty, often substantially. Thus, it may be the case that this wide variation represents the "reality" of current academic achievement of Texas students across grades and content areas.

- There was generally only a small change in central tendency (median or mean) of the panels' judgments across rounds. Seldom did the change in median cut across rounds exceed the standard error of measurement of the test. While individual panelists often changed their judgments substantially from round to round, the overall panels' recommendations typically changed little. Tabled below, for example, is the *median* raw-score change in the Met the Standard cuts from Round 1 to 2, from Round 2 to 3, and from Round 1 to 3 across the 26 English tests:

Median	**Number of Tests (of 26)**		
Raw Score Change	**Round 1 to 2**	**Round 2 to 3**	**Round 1 to 3**
Increase by 4 or more	3	2	6
Increase by 3	4	1	6
Increase by 2	6	2	3
Increase by 1	10	6	6
No change	**2**	**11**	**3**
Decrease by 1	1	1	1
Decrease by 2		1	
Decrease by 3 or more		2	1

Across grades and content areas, panels generally increased their Met the Standard cuts across rounds. The median raw-score increase from round to round across tests was 1 point. Commended cut changes from round to round were somewhat less variable than those for Met the Standard; the median Commended raw-score cut also increased by 1 point from Round 1 to Round 3. While it could not be argued that panelists were uninfluenced by the statewide projected data presented to them, these data typically appear to have had only limited impact on their recommended cut scores.

- As anticipated, there was a rather consistent decrease in the variability of judges' recommended cut scores across rounds. This was predictable given the literature on group decision-making processes and standard setting, which indicate increased levels of agreement among individual

panelists as group interactions continue. In the case of standard setting, a decrease in variability in judges' recommendations also typically occurs between Rounds 1 and 2 due to the improved understanding of the judgmental process—that is, the "mechanics" of making the recommendations. Across the 26 tests, most (though not all) showed reduced standard deviations of judges' recommendations from round to round—that is, increased agreement among the judges in the recommended cut scores.

Also as anticipated, the variability of judges' recommendations for the Commended cut scores was typically significantly lower than that for the Met the Standard cuts. This outcome is at least in part an artifact of the ceiling of the tests, which reduces the potential variability of judges' recommendations.

Appendix I presents a summary of the Evaluation Form results across all 36 tests. Data are presented separately by content area as well as being collapsed across all English tests. Differences across grades within content area were minor. Note that panelists who recommended standards for two tests completed two questionnaires with slightly different questions; this explains the sizable differences in sample sizes across individual questions. As these Evaluation Form data indicate, panelists overwhelmingly considered the major elements of the session to be adequate or better, were confident that the category descriptions used were reasonable and the judgments they made represented appropriate levels of student performance, that three rounds of ratings were sufficient, and that their opinions were treated with respect during the sessions. Differences in these data across panels—within and across content areas—were small; this is taken as an indirect indication of the comparability of the process and procedures across panels.

Appendix J provides a bar-graph summary of the anticipated statewide impact of the panels' recommendations by grade and content area. Readers can view the consistency of panel judgments both within and across content areas from these graphs. Given the fact that this process involved the collective judgments of almost 300 individuals assembled into 15 separate panels to recommend standards for 26 separate English TAKS tests, the amount of consistency across grades and content areas is remarkable.

Review of All Panel-Recommended Standards

After completion of all 15 initial sessions, four additional panels were convened. The purpose of these Review Panels, as they were termed, was to look at the recommended sets of cut-scores across all grade levels within content area for the 26 English TAKS tests. Thus, Review Panels were convened for English Language Arts (including Reading and Writing), Mathematics, Science, and Social Studies. A total of 50 panelists–all of whom were members of the earlier grade-based standard-setting

sessions—participated in the Review Panel meetings. Each panel met for a full day; a list of Review Panel members and the agenda for each of the sessions are presented in Appendix K.

The purpose of the Review Panel meetings was to provide an opportunity to inspect the recommended cut scores across all grade levels and to consider any revisions or adjustments that seemed advisable after looking across the several grades of data. It was decided to include representatives of the original grade-based panels in these discussions to permit them to express the views of their individual panels as any adjustments in their recommendations were considered. Clearly, it is a desirable outcome of any standard-setting activity that the outcome results in a reasonable and internally consistent set of standards across grade levels. For example, it would seem unreasonable for 65% of 3rd graders, 22% of 4th graders, 59% of 5th graders, and 62% of 6th graders to meet standard. In this hypothetical example, it would appear that the Grade 4 standards were set too high; that is, the Grades 3, 5, and 6 data are all rather consistent; and the Grade 4 percent is markedly out of line. It was the primary purpose of the Review Panels to ensure that the cuts were reasonably consistent across grade levels.

Two primary sets of data were used by the Review Panels to guide their deliberations. The first were the "impact data" bar charts shown in Appendix J. These permitted panelists to look across all grades and to gauge the consistency in the rigor of the earlier panels' standards. The second sets of data used by the panels were interim vertical scale score plots across grades generated for each content area. These interim values were calculated after the spring 2002 field test, the design for which permitted generation of a "growth-type" scale score system for each content area. The several interim scale scores presented for Review Panel consideration are shown in Appendix L.

Following a review of the above data, panelists discussed any inconsistencies in the cut scores across grades and explored how these inconsistencies could be reduced. All revisions in the original panels' recommendations were based on the actual test content. That is, before any revision in the interim cut scores was made, the Review Panel members reinspected the corresponding ordered test booklets and considered whether an adjustment could be justified in terms of the test content. As a general rule, panels sought to maintain the earlier panel recommendations; when revisions to those recommendations were made, the Review Panels attempted to keep revisions minimal and to be guided in any adjustments by the standard-error bounds for the tests and panel cuts (as shown for each test in Appendix H).

Across the 52 recommended cut scores, the Review Panels made no changes in the earlier panels' judgments in 32 cases. The table below summarizes the Review Panel changes (in raw score terms). Next to each change is indicated the test content area and grade for which this change was made. Note that each test (content area and grade) could be listed twice as two cuts apply to each test. As the table demonstrates, in almost all cases in which the Review Panels changed the earlier panels' cut scores, the change had the result of raising the cut scores. In general, Review Panels were of the opinion that the earlier grade-by-grade recommended standards were *slightly* too low.

This was especially true of the Science Review Panel, which raised the earlier recommendations for five of the six cuts, each by two or three raw score points.

Summary of Review Panel Changes of Earlier Panels' Recommendations

Amount of Change	No. of Cuts (of 52)	Test/Grade*
Raise cut by 6 pts.	1	W7
Raise cut by 5 pts.	1	SS8
Raise cut by 4 pts.	1	SS8
Raise cut by 3 pts.	3	M4,Sc11,Sc10
Raise cut by 2 pts.	5	R9,M3,Sc11,Sc10,Sc5
Raise cut by 1 pt.	8	R4,R3,M11,M11,M3,W4,W4,SS10
Make No Change	**32**	
Lower cut by 1 pt.		
Lower cut by 2 pts.		
Lower cut by 3 pts.	1	M10

* W=Writing, SS=Social Studies, M=Mathematics, Sc=Science, R=Reading

All Review Panel changes are listed on the Appendix H grade-by-grade summary-data pages. Of the 20 Review Panel changes, 12 applied to Met Standard cuts; 8 to Commended cuts. Three cuts each were revised for English Language Arts/Reading (of 18), for Writing (of 4), and for Social Studies (of 6); five cuts were revised for Science (of 6); and 6 (of 18) cuts were revised for Mathematics. The revisions had the general effect of "smoothing" the anticipated impact data across grades. The effect these changes had on the percent of students anticipated to score in each of the three categories is represented in bar-graph form in Appendix M. These graphs can be contrasted with the Appendix J bar graphs, which summarize the corresponding data prior to the Review Panels' activities.

In addition to making their data-based recommendations, the review panels generally discussed the operational aspects of the process of implementing the recommended standards. Across panels, two general recommendations to the SBOE emerged from these discussions. First, the panels recommended that the SBOE establish a "phase-in" period for the standards – that is, adopt the recommended standards, but provide a period of some small number of years during which these standards would become effective. Second, the panels recommended that the established standards be reviewed at some period of time, with this period of time ranging from the first time the tests "count" for a given cohort of students to two or three years after the standards are fully implemented. While there was some range in specifics for these two recommendations both within and across panels, all review panels made these two recommendations for SBOE consideration.

Following the Review Panel meetings, all data for the 15 "regular" standard-setting sessions and the four Review Panel sessions were summarized, checked for accuracy, and submitted to TEA as the final recommendations of the standard-setting panels. These recommendations in terms of raw score cuts are presented in Appendix H. Appendix H also indicates the anticipated statewide impact data for these cuts; the same data are portrayed graphically in Appendix M.

Convene Panels to Recommend Standards for the Spanish TAKS Tests

From October 14 through 23, panels were convened to recommend standards for the ten TAKS tests that in Spanish (Reading at Grades 3, 4, 5 and 6; Mathematics at Grades 3, 4, 5 and 6; Writing at Grade 4; and Science at Grade 5). The sessions for the Spanish TAKS tests were held after all English sessions were completed for two reasons. First, the Spanish tests were not field tested until September of 2002, so data were not available until the following month. In addition, it was decided during the planning of the sessions that panels recommending standards for the Spanish TAKS should be provided with the recommendations of the English TAKS tests for the corresponding grade and content area; these data were provided between the first and second rounds of the process to inform the Spanish panels as they made their judgments. Panelists were encouraged to consider these English cut scores as they made their recommendations for the Spanish TAKS tests.

A total of more than 75 panelists assisted with the six sessions held. As with the English tests, some of the Spanish TAKS panels recommended standards for tests at a pair of adjacent grades (Reading Grades 3 and 4, Reading Grades 5 and 6, Mathematics Grades 3 and 4, and Mathematics Grades 5 and 6); other panels considered only a single test (Writing Grade 4 and Science Grade 5). Appendix E includes the panelists for each session. Each panel was composed of 12 to 15 members. Panel members were nominated and selected using a process similar to that used to select panelists for the English TAKS tests.

All six sessions were facilitated by experienced contractor staff members; TEA staff members attended all sessions as observers and resources. Facilitators followed the same agendas, presentation sequence, overhead transparencies, and notes as were used for the English sessions. Panelists made three rounds of ratings as with the English tests. The only incremental information shared with the Spanish panels were the data concerning the cut scores recommended by the English Review Panels for the corresponding tests.

Results of the recommendations for the 10 Spanish TAKS tests are summarized in Appendix H. The projected statewide data in these cases are based on the Fall Study 2002 test data. Spanish panels were similarly deliberative to the English panels, making some amount of change from round to round. Spanish TAKS panels showed increasing levels of intra-panel agreement across rounds, reflecting again both

the comfort with the mechanics of standard setting and increased agreement as the group-discussion process continued.

The Evaluation Form results for the Spanish panels can also be found in Appendix I. Data are presented separately by content area as well as being collapsed across all Spanish tests. Differences across grades within content area were minor. These Evaluation Form data indicate the results were in general, similar to the English panels' results. For example, panelists overwhelmingly considered the major elements of the session to be adequate or better, were confident that the judgments they made represented appropriate levels of student performance, that three rounds of ratings were sufficient, and that their opinions were treated with respect during the sessions. Differences in these data across panels—within and across content areas—were small; this is taken as an indirect indication of the comparability of the process and procedures across panels.

The Round 3 recommendations of the Spanish TAKS panels paralleled quite closely those of the corresponding English panels. Of the 20 recommended cuts scores (2 per test times 10 tests), the Spanish panels recommended identical cuts to the English panels in 7 instances. In another 4 cases, the recommended Spanish cuts were 1 raw score point below the English cuts; in 7 cases, the recommended Spanish cuts were 2 points lower; and in 1 case the recommended Spanish cuts were 3 or 4 raw score points below those recommended for the English tests. Primarily because of the convergence of the English and Spanish panels' recommendations, it was not considered necessary to convene separate review panels for the Spanish tests. Appendix J also contains bar-graph summaries of the anticipated statewide impact of the panels' recommendations by grade and content area for the Spanish tests.