Lessons Learned: Decision Points for Evidence-Based Standard Setting

National Council on Measurement in Education
Philadelphia, Pennsylvania

Leslie Keng

Natasha Williams

Sonya Powers

April 2014

Abstract

This paper discusses the decision points that testing programs face as they implement the evidence-based standard-setting (EBSS) (O’Malley, Keng, & Miles, 2012) approach to establish performance standards for their assessments. As with any standard-setting process, EBSS requires multiple decisions to be made in order to implement the process in a way that produces reasonable results. Decisions may vary depending on the specific needs of the testing program. Lessons learned from implementations of the EBSS are synthesized and compared in four key areas. First, testing programs can select various types of empirical studies to include in the standard-setting process. Each implementation of the EBSS requires a rationale or procedure for selecting the empirical studies to use. Second, the standards to which the assessed curriculum is aligned and the process used to establish the performance level descriptors can cause slight differences in the way EBSS is implemented. Third, because EBSS includes policy considerations as part of the process, testing programs need to make decisions about how best to incorporate these policy considerations. And finally, decisions also need to be made regarding the type and amount of information shared with the standard-setting committee. Lessons learned from the different EBSS implementations provide helpful guidelines and practical advice to the field, especially as operational planning for standard setting begins for the common core assessments.

Keywords: standard setting, evidence-based, cut scores, large-scale assessment

Lessons Learned: Decision Points for Evidence-Based Standard Setting

Throughout any standard-setting process, there are multiple decisions that have to be made in order to implement the process in a way that produces reasonable results for the testing program. Evidence-based standard-setting (EBSS) (O'Malley, Keng, & Miles, 2012) also requires decision points throughout and those decisions may vary depending on the specific needs of the testing program.

This paper discusses the decisions made during various implementations of EBSS. The first section provides an introduction to the EBSS framework, followed by brief descriptions of different examples of EBSS implementations in various assessment programs. The subsequent sections focus on decision points and important considerations from the following key areas of an EBSS implementation: empirical studies, assessed curriculum, policy considerations, and standard-setting committees. Challenges faced and lessons learned from each of the example EBSS implementations are also described. The goal of this paper is to provide helpful guidelines and practical advice to the field, especially as operational planning for standard setting begins for the Common Core assessments and other testing programs.

Evidence-Based Standard Setting

The evidence-based standard-setting approach (EBSS; O'Malley, Keng, & Miles, 2012) was developed in response to the need to combine content-based judgments in traditional standard-setting methodologies (e.g., Angoff, 1971; Lewis, Mitzel, Green & Patz, 1999) with data from empirical studies that link performance on the assessment of interest to other related constructs or measures. For example, a testing program may require the cut scores on its assessments to indicate the degree to which a student is ready for postsecondary endeavors, such as college education and career opportunities. To inform the standard-setting process, research studies can be conducted that empirically link test scores on the assessment with performance on college and career readiness tests such as the SAT and ACT. Results of such empirical studies can then be used in a number of ways in the standard-setting process to help inform the content-based judgments made by the standard-setting committees.

Figure 1 provides a visual representation of the critical elements of the proposed EBSS approach. Each EBSS element is described further below.

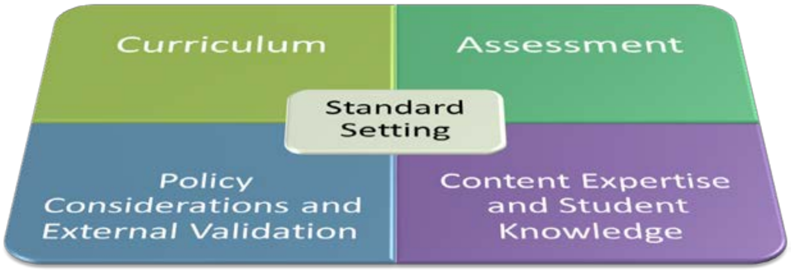

Curriculum
Assessment
Standard Setting
Policy Considerations and External Validation
Content Expertise and Student Knowledge

Figure 1: Critical Elements of the Evidence-Based Standard-Setting Approach

- Curriculum: The assessed curriculum and content standards for the testing program serve as the key underlying basis for several key components of the standard-setting process, including the general definitions for each performance level and grade or content area-specific performance level descriptors (PLDs).
- Assessment: Each test should be developed to assess the knowledge and skills described in the assessed curriculum and content standards for the specific grade level and content area and should adhere to the published blueprint and test specifications.

- Policy Considerations and External Validation: Results from research studies, which empirically associate test scores on the assessments of interest with performance on other related measures or external assessments, are used to inform the standard-setting process. Stakeholders and experts with experience in educational policy and knowledge of the testing program consider the study results when making recommendations about reasonable ranges for setting each performance standard.
- Expertise and Knowledge about Students and Subject Matter: Educators, including classroom teachers and curriculum specialists, bring content knowledge and classroom experience to the standard-setting process. They play an integral role in developing the PLDs and in recommending the performance standards.
- Standard Setting: Within the framework of EBSS, an established standard-setting method such as the bookmark method with external data (Ferrara, Lewis, Mercado, D’Brot, Barth, & Egan, 2011; Phillips, 2011) is used to recommend the cut scores for each assessment.

Steps in the EBSS Approach

Each implementation of the EBSS approach will differ depending on the unique circumstances and requirements of the testing program. The following five high-level steps, however, should constitute any EBSS implementation.

1. Conduct research studies
2. Develop general performance level definitions and specific PLDs
3. Develop reasonable ranges, or neighborhoods, for performance standards
4. Convene standard-setting committees
5. Review performance standards for reasonableness

A description of each high-level step in the EBSS approach is provided next.

Step 1: Conduct Research Studies

Extensive research should be conducted to support the standard-setting process. Studies can be conducted to empirically-link scores on tests in consecutive grade levels within the same content area (for example, grades 3 and 4 mathematics) for the assessment program. External validity studies that associate scores on assessments in the testing program to related external instruments or measures, such as the ACT and SAT, and other national or international assessments, may also be conducted.

Step 2: Develop General Performance Level Definitions and Specific PLDs

General performance level definitions describe the general level of knowledge and skills evident at each performance level across all grade levels and content areas. Specific PLDs (or simply, PLDs) are statements that articulate the specific knowledge and skills students typically demonstrate at each performance level of an assessment for a specific grade level and content area. They should provide a snapshot of students' academic characteristics based on performance on each assessment and reflect the breadth and depth of the content, skills, cognitive demand, and performance requirements evident in the assessed curriculum and content standards. The PLDs should be developed as an aligned system, describing a reasonable progression of skills within each content area.

A panel of stakeholders, including representatives from the department of education and their constituents, such as business leaders, policy experts, educators, and parents, can be convened to recommend the general performance level definitions; while committees of educators who are familiar with the expectations in the assessed curriculum and have experience with the students, can be convened to recommend the PLDs.

Step 3: Develop Reasonable Ranges for Performance Standards

The empirical studies conducted in Step 1 along with the general definitions and specific PLDs established in Step 2 can be used to derive ranges within which it would be reasonable to set each the performance standards on each assessment. These reasonable ranges are often referred to as *neighborhoods*.

Neighborhoods could be developed by policy and content experts within a department of education. Or, they could be established through the recommendation of a panel or committee comprised of stakeholders, many of whom may have helped recommend the general definitions or PLDs. The experts or panelists would consider policy implications of the performance standards along with the empirical study results to make neighborhood recommendations for the cut scores on each assessment.

Step 4: Convene Standard-Setting Committees

Committees consisting primarily of educators and subject matter experts use the general performance level definitions and PLDs (from Step 2), neighborhoods (from Step 3), and potentially select results from the empirical studies (from Step 1) to recommend performance standards, or cut scores, for each assessment. The committees are trained on and follow an established standard-setting method, such as the item mapping/bookmark method (Lewis, Mitzel, Green, & Patz, 1999) or the body of work method (Kingston, Kahl, Sweeney, & Bay, 2001) to make cut-score recommendations.

Step 5: Review Performance Standards for Reasonableness

After standard-setting committees have been convened to recommend performance standards for all of the assessments in the testing program, the cut-score recommendations are reviewed across grade levels and content areas to evaluate the reasonableness of the performance standards as a system and make policy-based adjustments as appropriate. This step can be done by the department of education, or by a panel or committee of stakeholders. The “briefing book” approach suggested by Haertel, Beimers and Miles (2012) can be adopted in which the recommended cut scores, research study results, neighborhoods, impact data and other relevant information (such as a proposed vertical scale) are summarized to help inform the reasonableness review process.

Example Implementations of EBSS

In recent years, there has been an increasing demand for K-12 testing programs to not only be a complete system in which performance across grade levels is linked vertically, but assessments at the high school level should also measure 21st century skills, which are indicators of college and career readiness. The EBSS approach, if properly planned and implemented, can yield performance standards that not only represent the students’ degree of mastery of the assessed curricula, but can also lead to cut scores that are vertically articulated across grade levels in each content area, yielding comprehensive assessment systems in which performance standards align and link back to elementary and middle school (grades 3–8) and project forward from high school to college and career readiness.

Example implementations of the EBSS approach include the standard-setting process for the following national or statewide testing programs:

- American Diploma Project (ADP)
- Common Core Aligned Assessments in Grades 3-8
- State of Texas Assessments of Academic Readiness (STAAR) Program
- Texas English Language Proficiency Assessment System (TELPAS)

Brief descriptions of how the EBSS approach was used each program are provided next.

ADP

A multistate collaborative effort took place from 2005 to 2007 where members of the ADP Network developed a common end-of-course (EOC) assessment in Algebra II. Because of the unique context and requirements of the assessment, the EBSS approach was used to establish performance standards on the ADP Algebra II test during the fall of 2007. Several types of empirical studies were conducted to link scores on ADP Algebra II to related measures at public schools as well as two- and four-year colleges. The briefing book method (Haertel, Beimers & Miles, 2012; Haertel, 1999) was then used to summarize results from the empirical studies as the primary input for standard-setting committee members in recommending cut scores for the ADP Algebra II test. More details about the standard-setting process for ADP Algebra II are described in O'Malley, Keng and Miles (2012).

Common Core Aligned Assessments in Grades 3–8

The EBSS approach was also used to establish performance standards for a system of Common Core aligned assessments in English language arts and mathematics from grades 3 through 8 for one of the member states. The charge for the state was not only for these assessments to measure the Common Core State Standards (CCSS), but for the performance standards to be rigorous, incorporate college readiness, and be vertically articulated across all grade levels within each content area. Additional details about the standard-setting process for this Common Core aligned assessment program can be found in Tong, Patterson, Swerdzewski & Shyer (2014).

STAAR

In 2012, the Texas Education Agency used the EBSS approach to establish performance standards on STAAR, the new K–12 statewide testing program. The EBSS approach was first used to set cut scores on the STAAR end-of-course (EOC) assessments given in high school. EBSS was then extended to the STAAR assessments at grades 3–8 by first setting the performance standards for middle school (grades 6–8), followed by elementary school (grades 3–5). The standards were set in this sequence to fulfill the state legislative mandate of anchoring the performance standards in college readiness for STAAR Algebra II and English III, which were the culminating STAAR assessments in the mathematics and English language arts content areas respectively. The standards set at the higher grade levels were then used to inform the cut scores at the lower grade levels for each content area thereby vertically articulating the performance standards from high school down through elementary school. Details of the standard-setting process for the STAAR EOC are described in LaSalle, Sedillo, Munoz, Ruff &

Phillips (2012), Keng, Murphy & Gaertner (2012), McClarty & Davis (2012), Williams, Keng & O'Malley (2012) and Texas Education Agency (2013). Boyd, Davis, Powers, Schwartz and Phan (2014) and Texas Education Agency (2013) detail the standard-setting process for the STAAR assessments in grades 3–8.

TELPAS

TELPAS is designed to assess the progress that English language learners (ELLs) in the state of Texas are making in acquiring the English language. TELPAS measures the state's second language acquisition curriculum standards, which support the ability of ELLs to acquire academic English while at the same time allowing them to engage meaningfully in regular, all-English academic instruction at their grade level. In compliance with Title III, Part A of the Elementary and Secondary Education Act (ESEA), TELPAS has been administered to Texas ELLs in grades K–12 in the language domains of listening, speaking, reading, and writing since the spring of 2008.

In 2013, a review of the proficiency level standards in the TELPAS reading domain was conducted. This standards review was predicated by the move to a new academic assessment program (STAAR) during 2012. STAAR was designed to be a more rigorous testing program, thus making it necessary to evaluate whether the standards in TELPAS reading could still be a meaningful indicator of the level of English language proficiency required to be successful on the STAAR reading assessments. The EBSS approach was used to review and make adjustments to the proficiency level standards on the TELPAS reading assessments in six grade levels/grade clusters, spanning grades 2 through 12. This represented the first use of the EBSS approach on a testing program designed to measure second language acquisition and proficiency. The use of this standard-setting approach proved useful not only in aligning the cut scores across all grade levels/grade clusters within the TELPAS reading domain, but also linking the proficiency level standards on TELPAS reading to performance on the appropriate STAAR reading assessments. More details about the standards review process for TELPAS reading can be found in Powers, Williams, Keng, & Starr (2014).

Clearly, each implementation of EBSS must take into account the unique requirements and challenges in the testing program of interest. In the remaining sections of this paper, key considerations and decision points for the above EBSS implementations are summarized and contrasted. The information is organized into four key areas of an EBSS implementation: empirical studies, assessed curriculum, policy considerations, and the standard-setting committees.

Empirical Studies

One of the distinctive features of the EBSS approach is the incorporation of empirical study results into the standard-setting process. In this section, five decision points for including empirical studies in standard setting are discussed: study selection, study classification, data collection design, analysis methodology and study usage. A summary of the specific decision made by each of the EBSS implementations is provided, followed by lessons learned from the various testing programs.

Study Selection

Each testing program and assessment has its own unique set of requirements and challenges, and will therefore need different types of empirical studies. Any testing program planning to conduct empirical studies, however, will quickly realize there is a large universe of potential studies from which it can choose. This fact, combined with the plethora of considerations (such as legislative requirements, data quality, timing, public perception, etc.), can make narrowing down and deciding on the specific set of studies to conduct a challenging endeavor. A systematic approach is therefore highly recommended.

Keng, Murphy and Gaertner (2012) describe a framework for selecting empirical studies that may prove useful for most testing programs. The framework includes three high-level steps.

1) *Identifying potential studies and study features* – A comprehensive list of all studies that could support the use or interpretation of the performance standards is generated. Each study is then rated on several key dimensions or features (such as data quality, the degree to which it fulfills legislative requirements, and the strength of the curricular relationship between the linked assessments) for comparison.
2) *Defining study selection guidelines* – Working with experts in curriculum, policy, and measurement, guidelines or rules for prioritizing and selecting potential empirical studies are created.
3) *Determining empirical studies* – By applying the study selection guidelines defined in step 2 to the list of potential studies and feature ratings generated in step 1, a set of empirical studies are chosen for use in the standard-setting process.

Details about the empirical study selection framework along with an example of how it was applied to an implementation of the EBSS approach can be found in Keng et al. (2012).

Study Classification

Once a set of empirical studies have been chosen, it is can be helpful to categorize the studies by the type of measures or instruments that are being empirically linked. O'Malley et al. (2012) provide one way of classifying various types of empirical studies. At a broad level, studies can be classified as *intra-program* studies or *external* studies. Intra-program studies empirically link measures or tests from within the testing program of interest. External studies, on the other hand, link tests from the testing program with measures or assessments from outside the program. Tables 1 and 2 summarize various types of intra-program and external studies that can be conducted to inform the standard-setting process.

Table 1. Intra-Program Studies (O'Malley, Keng & Miles, 2012)

Test-to-Test Linking Studies	
Description:	Examines the relationship between performance on related tests in the assessment program
Example:	Empirically link scores on grades 3 mathematics and grade 4 mathematics; or scores on the grade 8 mathematics and high school Algebra I end-of-course assessments
Test-to-Test Bridge Studies	
Description:	Evaluates the relationship between performances on a new test and a previous test in the assessment program
Example:	Empirically link scores on the new grade 5 reading test with the old grade 5 reading test that is being phased out
Test-to-Course Correlational Studies	
Description:	Looks at the relationship between test performance and course performance in the assessment program
Example:	Correlate scores on the grade 9 science test with high-school biology course grades, assuming that the biology curriculum is assessed on the grade 9 science test

Table 2. External Studies (O'Malley, Keng & Miles, 2012)

Concurrent Studies	
Description:	Match the performance on the test with performance on a related external assessment or measure typically taken at or around the same time
Examples:	Empirically link test scores on the grade 11 English language arts test with those on the SAT® Verbal test, assuming that the majority of students take the SAT® during their high school junior year
Predictive Studies	
Description:	Analyze the empirical relationship between performance on the test with performance on a related external assessment or measure taken at a different time (either before or after)
Example:	Project scores on the grade 7 mathematics test to scores on the EXPLORE® Math test; or project scores on the EXPLORE® Science test to those on the high-school physics test. The EXPLORE® test is usually taken by students in 8th grade
Cross-Sectional Studies	
Description:	Evaluate the empirical relationship between test performance and some external criteria or definition by administering the test to a group of test-takers that satisfy the criteria and a group of test-takers that do not
Example:	To inform the setting of a college-readiness performance standard on the high school Algebra II test, administer the Algebra II test to college freshmen enrolled in an entry-level college algebra course at the start of the semester, then compare the test performance of those who pass the course (i.e., college ready) and those who do not (i.e., not college ready)
Longitudinal Studies	
Description:	Examine the relationship between test performance and some external measures or criteria by tracking examinees on the measure or criteria across time
Example:	To inform the setting of a college-readiness performance standard on the high school Algebra II test, track the students who took the Algebra II test in high school over time by obtaining their college grade point average (GPA) at the end of each of their college years

As noted by O'Malley et al. (2012), Tables 1 and 2 do not represent exhaustive lists of the types of empirical studies that can be conducted. Each testing program should consider its needs and requirements and decide on the types of studies to conduct. These decisions would in turn help to determine what data collection design and analysis methodology to implement, both of which are discussed next.

Data Collection Design

Decisions also need to be made about how to collect the data for each study. A substantial amount of planning is often required to consider issues such as the data source (From where is the data coming?), time constraints ("By when do we need the data?"), variables of interest ("What types of measures and characteristics are needed?"), and level of data ("Do we need student- or campus-level test scores?").

If a study involves linking performance on two different measures (e.g., test scores on a statewide reading assessment and scores on the ACT Mathematics test), then additional considerations, such as when each test is typically administered (e.g., once a year during the spring vs. multiple times throughout the school year), what type of students generally take each assessment (e.g., college-bound juniors vs. all students), and what variables can be used to merge the data sets (e.g., unique student identifier, first/last name, date of birth etc.), must also be taken into account. Ideally, a single group design, where all of the data are collected from a single group of students, is desirable when empirically linking two measures. However, this may be logistically challenging in some test administration scenarios. In those cases, sample matching methodologies such as coarsened exact matching (Iacus, King, & Porro, 2011) or propensity score matching (Rosenbaum & Rubin, 1983) may be used to simulate a single group of students who have scores on both measures.

Analysis Methodology

In general, two statistical methods have been used to analyze the data collected from single- or matched-group designs: *equipercentile linking* and *regression-based projection*.

With the equipercentile linking method, the score distributions on each of the measures are computed and scores with equivalent percentile ranks are empirically linked across the two measures. This method has been used to link scores on the SAT and ACT assessments (Dorans, Lyu, Pommerich, & Houston, 1997; Pommerich, Hanson, Harris & Sconing, 2004). The method is appropriate in cases where the curricula assessed by the two measures or instruments are similar enough such that it can be assumed that they are measuring the same underlying construct. Results of the equipercentile linking method are usually summarized in concordance tables, which map the scores on one assessment to their equivalent scores on the other assessment. An example of a concordance table between the SAT and ACT scores can be found at the ACT web site (http://www.act.org/solutions/college-career-readiness/compare-act-sat/).

Regression-based linking is useful for empirical studies in which no assumptions about score equivalency can be made between the two measures or instruments being linked. Correlations and scatterplots should be generated first to evaluate the appropriateness and strength of the linear regression relationship. Two types of regression-based approaches have been used: logistic regression and ordinary least square (OLS) regression. Logistic regression estimates the probability that an examinee would achieve a certain level of performance on one assessment, given their performance on the other assessment. OLS regression estimates the expected score an examinee would obtain on one assessment, conditional on their performance on the other assessment. The choice of which regression-based approach to utilize depends on various factors, such as the type of data available for each measure, and the intended interpretation or use of the study results. Results from regression-based linking can be presented in expectancy tables (for logistic regression) or projection tables (for OLS regression). These tables look similar in form to the concordance tables that are yielded from equipercentile linking. One important distinction, however, is that the mapping of scores between the two assessments in a concordance table is symmetric (because they are considered equivalent), but that is not the case for the scores in expectancy and projection tables (as linear regression in not symmetric). As such, careful consideration should be taken in deciding on the direction of the regression relationship.

Please refer to Keng et al. (2012) for a more complete discussion of these analysis methods, including technical details and example results.

Study Usage

Empirical study results may be used to inform three aspects of the standard-setting process: (1) Empirical study results can be a key source of information for developing reasonable ranges (or neighborhoods) within which the various cut scores can be set (i.e., Step 3 of the EBSS approach). Depending on the purpose and data quality of each study, the results may help inform the boundaries (lower or upper bound) of the neighborhoods, or help in evaluating whether the proposed neighborhoods are reasonable; (2) Results from the studies may also be provided to the standard-setting committees after each round of judgment to help provide context or as a reality check for their recommended cut scores (i.e., Step 4 of the EBSS approach); (3) After all standard-setting committees have recommended cut scores, the study results can be used to evaluate the reasonableness of the performance standards as a system (i.e., Step 5 of the EBSS approach).

Summary of Implementations

Table 3 provides a summary of the empirical studies conducted as part of each EBSS implementation and how the studies where used in the standard-setting process.

Table 3. Summary of Empirical Studies for EBSS Implementations

EBSS Implementation	ADP	Common Core	STAAR	TELPAS
Study Types				
Test-to-Test Linking Studies			√	
Test-to-Test Bridge Studies			√	√
Test-to-Course Correlational Studies			√	
Concurrent Studies	√		√	√
Predictive Studies	√	√	√	√
Cross-Sectional Studies	√	√	√	√
Longitudinal Studies				
Study Usage				
Reasonable Range Development		√	√	√
Standard-Setting Committee Feedback	√	√	√	√
Cut Score Reasonableness Review			√	√

Lessons Learned

Advanced Planning

The importance of setting aside sufficient time to plan, obtain data and conduct the empirical studies cannot be over-emphasized. The process of selecting the set of studies to conduct is an iterative process that may involve several rounds of discussion with various stakeholders and experts in curriculum, policy and measurement. Securing data sharing agreements from the owners of the assessment instruments and examinee records can also be time-consuming. The temporal relationship between the measures being linked is also a key consideration. For example, if the goal of a study is to examine the predictive relationship between scores on an EOC assessment typically taken by high school juniors or seniors and first-year college grade point average (GPA), then there would be a one- to two-year gap between a student's score of the EOC assessment and the college GPA. If a single-group design is desired for this study, then data collection needs to begin at least two years prior to when performance standards are needed for the EOC assessment. This was the case for the performance standards on the STAAR EOC assessments. Performance standards needed to be in place by the spring 2012 administration. As such, data collection for the EOC assessments began as early as 2009 (Keng et al., 2012).

Use of Vertical Scale

If the assessments on which performance standards need to be set are related through a vertical scale, then it is recommended that the vertical scale study be constructed in advance so that it can be used to inform neighborhood development and to evaluate the reasonableness of the recommended cut scores. In other words, the vertical scale can be used as another type of “empirical study” in the standard-setting process. A vertical scale was available and played an important role in the standard-setting process for the STAAR assessments in grades 3-8 and the TELPAS program (Boyd et al., 2014; Powers et al., 2014).

Impact of Motivation and Opportunity to Learn

In many cases, performance standards are needed on a test before its initial high-stakes administration. Assessment data from the testing program used in the empirical studies must therefore be collected prior to this initial administration. Consequently, this data may be based on the performance of students who have not been fully instructed on the assessed curriculum, or who are not motivated to try their best on the test because they realize it does not “count”. This may have a negative impact on the validity of the study results as it introduces statistical error due to construct irrelevant variance. In some cases, it may be possible to identify student records that are severely impacted by motivation (for example, students who did not attempt a substantial number of test items or left a constructed responses item such as an essay completely blank). Such records can be filtered out before the data are analyzed. In other cases, it may be possible to use historical performance data (for example, from when a new testing program was previously introduced) to estimate a statistical adjustment due to motivation and opportunity to learn. The fact is, however, that because of the unique set of factors and circumstances faced by each testing program, finding or conducting research to support the use of motivation and opportunity to learn adjustments is not a trivial undertaking and is prone to measurement error. As such, it may be best to acknowledge the limitations of such studies when sharing the results and using them to inform the standard-setting process.

Longitudinal Studies

It would be ideal if longitudinal studies, which analyze the performance of test takers on the tests and various measures of interest across time, could be conducted and used in the standard-setting process. However, this is usually not feasible due to the practical limitation of when performance standards need to be in place for the assessment of interest. As shown in Table 3, none of the EBSS implementations were able to incorporate longitudinal studies into the standard-setting process. It is still recommended that the testing program track the performance of its test takers across time so that longitudinal studies may eventually be conducted and used as part of a standards review or validation process several years after the initial standard-setting process.

Assessed Curriculum

The type of curriculum framework or content standards (e.g., college and career readiness, English language proficiency) to which the assessed curriculum is aligned and the process used to establish the performance level descriptors (PLDs) may lead to slight differences in how the EBSS approach is implemented. In this section, we describe the roles that assessed curriculum played in the various EBSS implementations in terms of content analysis and PLD development.

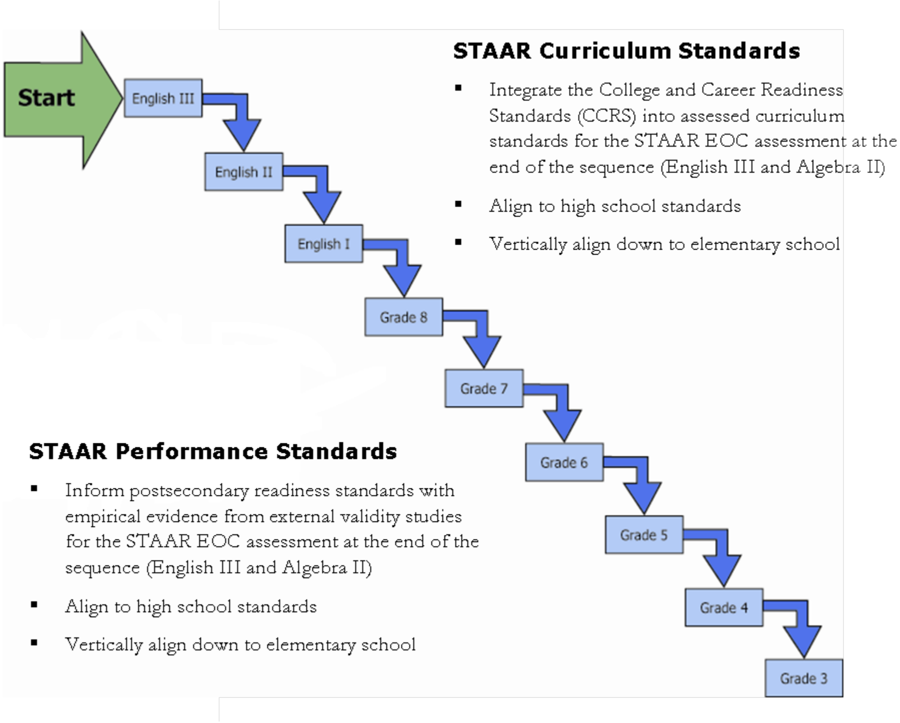
Content Analysis

As with traditional standard-setting methods, subject matter experts (SMEs) such as educators and curriculum specialists also play important roles in the EBSS approach. Prior to participating in the standard-setting committees (discussed in a later section), SMEs should be involved early and often in the standard-setting process. For the various EBSS implementations, SMEs helped conduct numerous content-based analyses that were instrumental to supporting the content validity of the performance standards and the testing program in general.

Alignment

In the context of education, alignment is defined as the extent to which various components, such as curriculum, instruction, and assessments, fit together as a system to help accomplish the goal of student learning (Webb, 1997). When a testing program includes a system of assessments (for example, tests in reading and mathematics from elementary through high school), the vertical alignment of performance standards across content-area assessments is important in supporting the predictive validity of the test scores. One key prerequisite of achieving vertical alignment of performance standards is to ensure that the curriculum standards assessed by the tests are vertically aligned. The task of curriculum alignment can be a large undertaking and usually requires significant time commitment from many groups of SMEs involved in the testing program.

An example of an EBSS implementation that required vertical alignment of its standards is the STAAR program in Texas. State legislation not only required vertical alignment of the grade-to-grade performance standards of the STAAR assessments within a content area, but the performance standards for the culminating high school assessments in English III and Algebra II also needed to be predictive of each examinee’s readiness for postsecondary endeavors, such as college and careers. To fulfill this requirement, content specialists from the state’s education agency worked with committees of SMEs from across the state to establish and refine the assessed curricula such that it was vertically aligned within content-area STAAR assessments from elementary to high school. In addition, the state worked with SMEs from higher education to integrate the state's College and Career Readiness Standards (CCRS) into the assessed curricula for STAAR English III and Algebra II. This extensive and comprehensive curriculum alignment work commenced years before the STAAR program was launched and helped lay the ground work for the vertical alignment of the STAAR performance standards.



STAAR Curriculum Standards

- Integrate the College and Career Readiness Standards (CCRS) into assessed curriculum standards for the STAAR EOC assessment at the end of the sequence (English III and Algebra II)
- Align to high school standards
- Vertically align down to elementary school

STAAR Performance Standards

- Inform postsecondary readiness standards with empirical evidence from external validity studies for the STAAR EOC assessment at the end of the sequence (English III and Algebra II)
- Align to high school standards
- Vertically align down to elementary school

Figure 2. Vertical alignment of curriculum and performance standards in STAAR

Figure 2 provides a visual illustration of the alignment of curriculum and performance standards in the STAAR English language arts/reading content area. Specific details about the curriculum alignment work conducted for the STAAR program can be found in LaSalle et al. (2012).

Gap Analysis

When one of the goals of a testing program is to increase the rigor of the curriculum assessed by the tests along with their performance standards, then it is usually informative to elicit the help of SMEs to conduct gap analyses to compare the old and new assessed curricula. LaSalle et al. (2012) provides an example of the gap analysis done to compare the assessed curricula of the old program and the new STAAR program. Another example of a gap analysis was the text complexity study described in Powers et al. (2014). In this study, reading passages on the TELPAS assessments and the STAAR assessments were compared on four complexity criteria, including purpose and meaning, organization and structure, language, and knowledge demands.

Curriculum Overlap

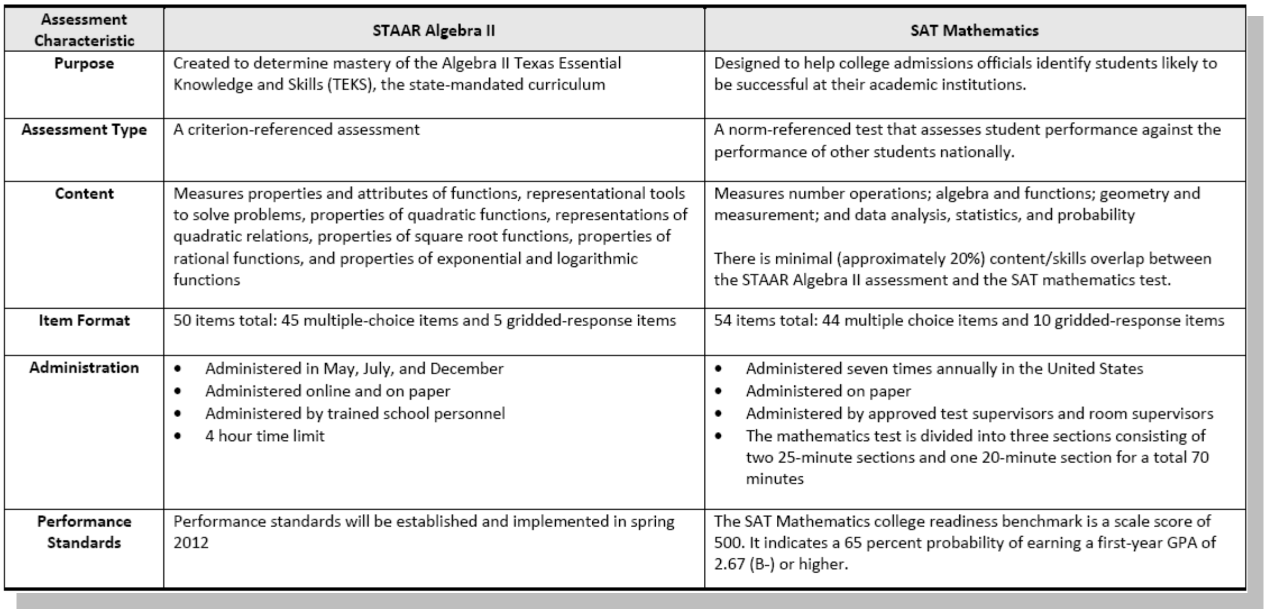
To help evaluate whether it is reasonable to empirically link two measures, it is important to look at the degree of similarity or overlap between the curricula assessed by the two instruments. SMEs can play a critical role in conducting these curriculum overlap analyses. To make the analysis task manageable for the SMEs, it is important to specify characteristics of each instrument that should be compared. It can also be helpful to develop a metric for classifying the degree of overlap.

As an example, for all external validity studies conducted for the STAAR EOC standard-setting process, each pair of tests was compared on six assessment characteristics: purpose, assessment type, administration, item formats, testing time and performance standards. An overall rating of the degree of content overlap was also assigned using the classification rules included in Table 4.

Table 4. Example Content Overlap Classification Rules

Overlap Rating	Classification Rules
0 – No Content/Skills Overlap	Different content areas assessed by tests
1 – No Content/Skills Overlap	Same content areas but 0% overlap in assessed student expectations (SEs) and test blueprints
2 – Minimal Content/Skills Overlap	Same content areas with 1–25% overlap in assessed SEs and test blueprints
3 – Some Content/Skills Overlap	Same content areas with 26–50% overlap in assessed SEs and test blueprints
4 – Moderate Content/Skills Overlap	Same content areas with 51–75% overlap in assessed SEs and test blueprints
5 – Strong Content/Skills Overlap	Same content areas with 76–100% overlap in assessed SEs and test blueprints

Figure 3 shows the summary of the content overlap analysis for the STAAR Algebra II-to-SAT mathematics study.



Assessment Characteristic	STAAR Algebra II	SAT Mathematics
Purpose	Created to determine mastery of the Algebra II Texas Essential Knowledge and Skills (TEKS), the state-mandated curriculum	Designed to help college admissions officials identify students likely to be successful at their academic institutions.
Assessment Type	A criterion-referenced assessment	A norm-referenced test that assesses student performance against the performance of other students nationally.
Content	Measures properties and attributes of functions, representational tools to solve problems, properties of quadratic functions, representations of quadratic relations, properties of square root functions, properties of rational functions, and properties of exponential and logarithmic functions	Measures number operations; algebra and functions; geometry and measurement; and data analysis, statistics, and probability There is minimal (approximately 20%) content/skills overlap between the STAAR Algebra II assessment and the SAT mathematics test.
Item Format	50 items total: 45 multiple-choice items and 5 gridded-response items	54 items total: 44 multiple choice items and 10 gridded-response items
Administration	• Administered in May, July, and December • Administered online and on paper • Administered by trained school personnel • 4 hour time limit	• Administered seven times annually in the United States • Administered on paper • Administered by approved test supervisors and room supervisors • The mathematics test is divided into three sections consisting of two 25-minute sections and one 20-minute section for a total 70 minutes
Performance Standards	Performance standards will be established and implemented in spring 2012	The SAT Mathematics college readiness benchmark is a scale score of 500. It indicates a 65 percent probability of earning a first-year GPA of 2.67 (B-) or higher.

Figure 3. Example Content Overlap Summary for STAAR Algebra II-to-SAT Mathematics

Based on the comparison of the assessed SEs and test blueprints of the two tests, the overall content overlap was rated as strong (5) for this particular study. More details about the content overlap analysis process can be found in LaSalle et al. (2012).

PLD Development

PLDs are an essential component of any standard-setting process as they articulate the specific knowledge and skills that examinees typically demonstrate at each performance level of a test. Because of their domain expertise and experience with test takers, SMEs play an integral role in developing the PLDs. Each testing program needs to make decisions about the PLD development process based its own unique set of timelines, resources and requirements. Specific decision points can include:

- *Who* should develop the PLDs?
- *When* should the PLDs be developed?
- *How* should the PLDs be developed?

Table 5 provides a summary of the decision made by each of the EBSS implementations.

Table 5. Summary of Empirical Studies for EBSS Implementations

Decision Point	ADP	Common Core	STAAR	TELPAS
Who?	Postsecondary educator committee Consortium SMEs	PLD committees Standard-setting committees	PLD committees State education agency SMEs	SMEs from the field and state education agency
When?	Prior to standard-setting committee	Prior and during standard-setting committees	Prior to standard-setting committees	Prior to previous standard setting
How?	Developed as part of judgment studies and included in briefing book for standard-setting committee	Initial development by PLD committees Refined and finalized by standard-setting committees	Developed by PLD committees and state agency SMEs	Developed based on the state's English language proficiency standards in statue

Lessons Learned

As one can glean from the types of content-related work conducted for the various EBSS implementation, SMEs play a vital role in the standard-setting process. Because most SMEs are neither measurement not policy experts, it is important to communicate expectations and requirements in a clear and structured manner to SMEs to help avoid confusion and redundancy caused by ambiguity. For example, if a gap analysis needs to be conducted between the old and new curriculum, it would be well-worth the time to work with the SMEs to derive a detailed rubric or rating scheme, clearly outlining the expectations in each category, prior to any analysis work. It may be helpful to work through one example analysis as a group to make sure that everyone is on the same page.

Ample time should be allotted to conduct the content analysis and develop the PLDs. These processes can be very time consuming, especially if the work involves coordination of

SMEs with the test vendor, at the state agency, from within a state, or even across multiple states. Set the expectation with the SMEs that this process is an iterative one that could involve several rounds of discussion and revisions, possibly with different types of stakeholders.

Finally, it can be beneficial to the efficiency and quality of the content analysis and PLD development processes if the policy experts and measurement specialists take time to acquire a deeper understanding on the assessed curriculum so that they are speaking the same “language” as the SMEs.

Policy Considerations

Another unique characteristic of the EBSS approach is the intentional incorporation of policy considerations into the standard-setting process. The different types of stakeholders with policy expertise that should be involved in the process should be considered. The roles that the policy experts play in the standard-setting processes should also be defined. Careful thought should also be put into how the results of empirical studies and curriculum analysis are communicated to the policy experts so that their contribution to the standard-setting process is well-informed. This section discusses these three areas of policy considerations for an EBSS implementation.

Types of Policy Experts

Unlike SMEs and measurement specialists, the types of policy experts that may be involved in any given standard-setting process can vary greatly in terms of assessment background, knowledge of the testing program, and their interests or stakes in the outcome of the standard setting. The requirements of a testing program – both legally and to satisfy the court of public opinion – are unique. Thus, the group of policy experts is expected to differ across programs and may even change across time for a particular program. As an example, Table 6 provides a list of the types of policy experts that were recruited for the STAAR standard-setting process, along with the rationale for including each type (Texas Education Agency, 2013).

Table 6. Policy Experts Considered for the STAAR Standard-Setting Process

Recruitment Group	Rationale
Business/workplace leaders	Career/workforce readiness is one of the stated goals of the STAAR assessment system.
Higher education representatives	College readiness is one of the stated goals of the STAAR assessment system.
Legislative staffers	Can provide information about legislative intent behind the requirements for STAAR.
Policy experts	Can offer policy expertise related to postsecondary readiness at the state and national level.
Texas educators/educators with policy experience	This group may include teachers and administrators, such as principals, curriculum specialists, and superintendents. The former can bring content knowledge and classroom experience, while the latter can offer specific knowledge about how test results are used at the district, campus, and classroom levels.
Special population representatives	Represent the perspectives of English language learners and students served by special education.
Community representatives	Represent the interests of other stakeholders, such as PTA representatives.

Role of Policy Experts

Policy experts can be involved in several steps in the standard-setting process, providing valuable input into the meaning and location of the performance standards. The roles that policy experts have played in the EBSS implementations are described next.

Generate Policy Labels and Definitions

Policy experts can help generate labels for the performance categories (e.g., basic, proficient and advanced) and craft general definitions of what students in each category are expected to know and do. These general definitions, sometimes referred to as the general or policy-level performance level descriptors (PLDs), apply to the tests at all grade levels and content areas in the testing program and are usually the basis for the specific PLDs. For example, policy experts from the state department, with help from administrators in the field, shaped the labels and policy-level PLDs for the four achievement levels (Levels 1-4) on the Common Core aligned assessments in grades 3-8 described in Tong et al. (2014). A Performance Descriptor Advisory Committee (PDAC) comprised of a diverse group of stakeholders from public education and higher education in Texas was convened to recommend to the Commissioner of Education the performance labels (Unsatisfactory Academic Performance, Satisfactory Academic Performance, and Advanced Academic Performance) and policy definitions for the STAAR assessments (O'Malley et al., 2012; Texas Education Agency, 2013).

Recommend Neighborhoods

Because they understand or are impacted by the intended or unintended consequences of the performance standards, policy experts can play a key role in helping to locate the neighborhoods, or ranges in which it would be reasonable for the cut scores to be, for each test. This range-finding process takes place before the standard-setting committee convenes and the resulting neighborhoods are key input into the standard-setting committee's work. It should be noted that the concept of neighborhoods is not one that is unique to the EBSS approach. The idea of deriving a reasonable range for cut scores to inform the standard-setting committee was also implemented in the other benchmark standard-setting methods (Ferrara, Lewis, Mercado, Egan, D'Brot, & Barth, 2011; Phillips, 2012). What EBSS does is to bring together the experience of the policy experts and results from empirical studies to develop the neighborhoods.

For the STAAR program, this was accomplished by convening a policy committee consisting of the stakeholder groups shown in Table 6 prior to bringing in the standard-setting committee (McClarty & Davis, 2012; TEA, 2013). Similarly, a pre-policy meeting was held to establish reasonable ranges for the Common Core aligned assessments in grades 3-8 (Tong et al., 2014).

Review the Recommended Cut Scores for Reasonableness

After the standard-setting committees have made their cut-score recommendations, policy experts can also be consulted with on the reasonableness of the recommended standards. The goal of this reasonableness review process is to look at the performance standards as a system and evaluate whether they are supported by empirical study results, in compliance with legislative requirements, and in alignment with policy expectations and public perceptions. Adjustments to the cut scores can occur in this process. Some form of post-standard-setting reasonableness review by policy experts was implemented for all four EBSS implementations described in this paper. In most cases, the performance standards coming out of the reasonableness review process were sent forward to the rule-making authority for final approval and then implemented in the testing program.

Communicating Information to Policy Experts

To help policy experts engage in the standard-setting process, it is crucial that information such as content analysis and empirical study results be presented to them in an accurate, concise, yet understandable manner. This can be a daunting undertaking because the volume of information to communicate is usually substantial, especially if there are a large number of studies. All EBSS implementations presented in this paper grappled with the challenge of data reduction and visualization. Details about the approaches taken for the various implementations can be found in the respective papers. In this subsection, three specific approaches for communicating analysis and study results are highlighted.

Study Profiles and Data Quality Summary

Study profiles provide a snapshot of the salient pieces of an empirical study that a policy expert needs to know to make an informed decision. For the STAAR EOC standard-setting process, a study profile was produced for every external validity study. Each profile included sections on the sample attributes, statistical properties, degree of content overlap, and key characteristics of the assessments in each study. Appendix 1 shows an example of the study profile created and presented to the policy committee.

Designed as a complement to the study profile, the data quality summary takes the ratings given to each empirical study in its study profile and summarizes them into a single easy-to-read document. For example, for the STAAR EOC process, each study was rated on five dimensions: motivation, representativeness, sample size, correlations, and content overlap. An overall rating, computed as a weighted average of the five dimension ratings, was also computed. The data quality summary that was generated and shown to the policy committee for the STAAR EOC standard-setting process is shown in Appendix 2.

The complete set of STAAR EOC external validity study profiles along with the data quality summary can be found at: http://www.tea.state.tx.us/staar/vldstd.aspx.

Empirical Number Lines

An empirical number line is a visual way to illustrate how various empirical study results fall relative to one another. It can be an informative tool to help policy experts make recommendations about cut-score neighborhoods. Figure 4 shows an example of an empirical number line for the STAAR EOC standard-setting process.

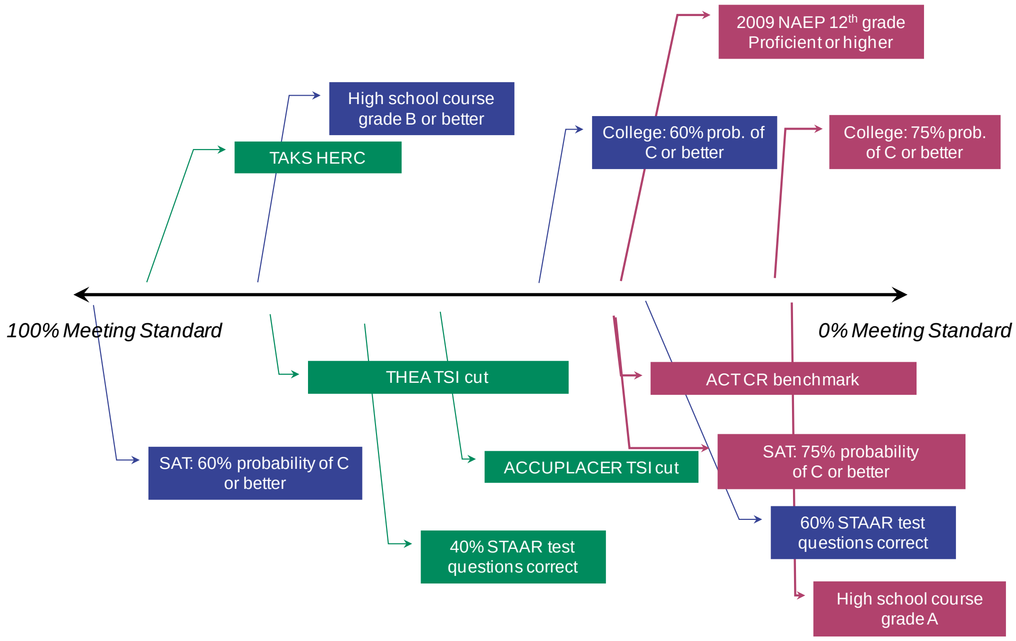
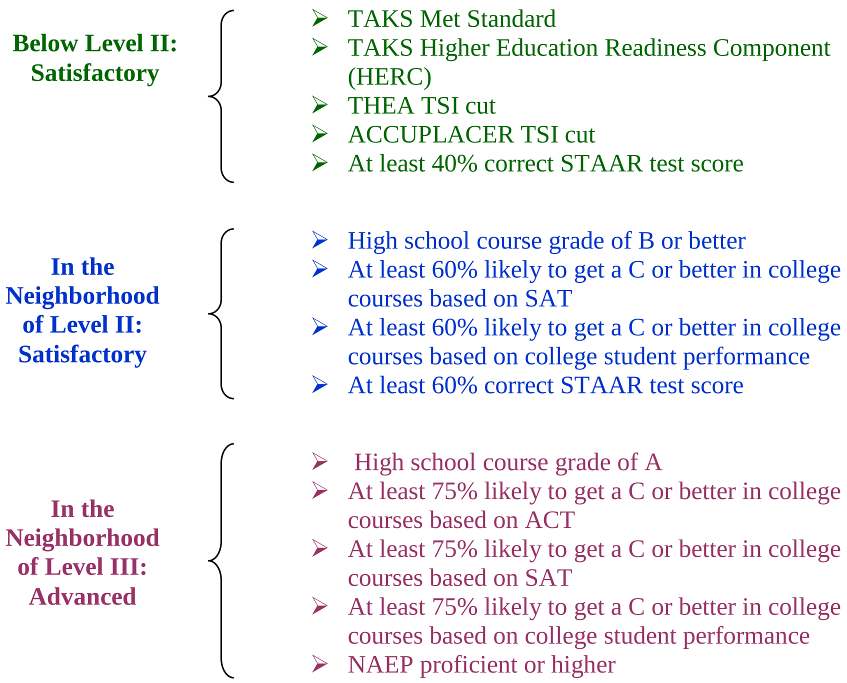

2009 NAEP 12th grade Proficient or higher
High school course grade B or better
TAKS HERC
College: 60% prob. of C or better
College: 75% prob. of C or better
100% Meeting Standard
0% Meeting Standard
THEA TSI cut
ACT CR benchmark
SAT: 60% probability of C or better
ACCUPLACER TSI cut
SAT: 75% probability of C or better
60% STAAR test questions correct
40% STAAR test questions correct
High school course grade A

Figure 4. *Example Empirical Number Line for a STAAR EOC Assessment*

To help policy experts engage with the empirical number lines, it may also be helpful to develop guidelines for which studies fall into the different cut-score neighborhoods for an assessment. Figure 5 shows the neighborhood development guidelines for the STAAR EOC standard-setting process. Note how the color scheme in these guidelines matches the color scheme in the empirical number line in Figure 4.



Below Level II: Satisfactory

- TAKS Met Standard
- TAKS Higher Education Readiness Component (HERC)
- THEA TSI cut
- ACCUPLACER TSI cut
- At least 40% correct STAAR test score

In the Neighborhood of Level II: Satisfactory

- High school course grade of B or better
- At least 60% likely to get a C or better in college courses based on SAT
- At least 60% likely to get a C or better in college courses based on college student performance
- At least 60% correct STAAR test score

In the Neighborhood of Level III: Advanced

- High school course grade of A
- At least 75% likely to get a C or better in college courses based on ACT
- At least 75% likely to get a C or better in college courses based on SAT
- At least 75% likely to get a C or better in college courses based on college student performance
- NAEP proficient or higher

Figure 5. Neighborhood Development Guidelines for a STAAR EOC Assessment

Using these two tools, the policy committee was able to make cut-score neighborhood recommendations for the STAAR EOC assessments. McClarty and Davis (2012) and Texas Education Agency (2013) provide more details of how these tools were used in the standard-setting process.

Validity Studies Crosswalk

The validity studies crosswalk is another way of showing how the empirical study results fall relative to one another. However, rather than providing a visual representation, the crosswalk shows the study results in relation to a specific test form. Appendix 3 shows the validity studies crosswalk created for the ADP Algebra II standard-setting process. This crosswalk was part of the briefing book shown to policy experts and the standard-setting committee. Details about how this crosswalk was constructed and used can be found in O'Malley et al. (2012).

Lessons Learned

The approaches for involving policy experts in the standard-setting process can be as varied as the requirements of each testing program. Table 7 shows how policy experts contributed to the standard-setting processes for the four EBSS implementations.

Table 7. Policy Experts Involvement in EBSS Implementations

EBSS Implementation	Steps in Standard-Setting Process
ADP Algebra II	Judgment Studies Standard-Setting Meeting
Common Core Aligned	Policy-Level PLDs Standard-Setting Meetings Reasonableness Review
STAAR	Performance Labels and Policy Definitions Neighborhood Development Standard-Setting Meetings Reasonableness Review
TELPAS	PLD Development

One important takeaway is therefore to keep an open mind about the role of policy experts in the standard-setting process. Different stakeholders and policy groups invariably have different interests and agendas when it comes to the testing program and its performance standards. Understanding how to satisfy each party's interests, without marginalizing another's, can often be more of an art than a science. Plan for an ample amount of time to discuss and decide which types of policy experts to include in the process, from which organizations or stakeholder groups the experts should come, and which steps in the process to include each expert. Policy experts usually have a plethora of competing priorities and commitments. Thus, scheduling any committee meetings well in advance is highly recommended.

For the proceedings of any committees comprised primarily of policy experts, it may be advisable to invite independent third-party (i.e., not from the state agency or testing vendor) facilitators with experience and credentials in the field of education. This can help avoid any perceived or actual conflict of interest with the committee members while lending credence to the process. These experienced facilitators can also serve as advisors and advocates for the standard-setting process and the testing program.

Finally, deciding on what information to present to policy experts and how to effectively communicate it can take a substantial amount of time and effort, usually involving several iterations of trial and error. The examples shown in this section are exemplars that served the needs of the EBSS implementations described, but they are by no means the only approaches to summarizing and articulating study results for the purpose of standard setting. Indeed, it may be beneficial to stay on top of the latest research and innovations in reporting and data visualization to help inform or refine the current approaches. Feedback from policy experts (and non-technical audiences in general) through surveys or focus groups can also provide valuable input into the effectiveness of various reporting and data visualization techniques and help advance these important areas of research.

Standard-Setting Committee

When convening the standard-setting committee, there are many decisions that testing programs must make based their constraints and requirements. In this section, two specific decision points that an EBSS implementation should consider are discussed: the standard-setting method that the committees follow to recommend the standards, and the feedback data provided to the committee members after each judgment round.

Standard Setting Method

Several traditional standard-setting methods can be used by the standard-setting committees in an EBSS implementation. Table 8 shows the specific standard-setting methods employed by the four EBSS implementations described in this paper.

Table 8. Standard-Setting Committee Methods in EBSS Implementations

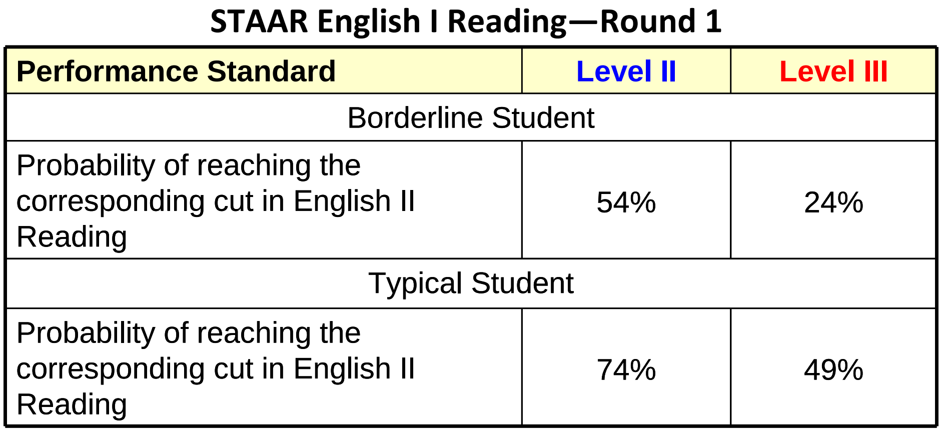
EBSS Implementation	Standard-Setting Committee Method
ADP Algebra II	Briefing Book (Haertel, Biemers & Miles, 2012; Haertel, 1999)
Common Core Aligned	Bookmark/Item-Mapping (Lewis, Mitzel, Green & Patz, 1999)
STAAR	Bookmark/Item-Mapping (Lewis, Mitzel, Green & Patz, 1999) Yes/No Procedure (Impara & Plake, 1997)
TELPAS	Bookmark/Item-Mapping (Lewis, Mitzel, Green & Patz, 1999)

While future implementation could consider using other traditional standard-setting methods, it should be noted that not all methods work well with the concept of cut-score neighborhoods. It would be challenging, for example, to figure out how to constrain the item probability judgments of standard-setters in an Angoff approach (Angoff, 1971) so that the resulting cut-score recommendations falls within the pre-determined reasonable ranges for each of the performance standards. As one may deduce from Table 8, the bookmark/item-mapping procedure with its use of ordered item booklets (OIBs) seems to lend itself nicely to cut-score neighborhoods. The interested reader may refer to the detailed description in Williams et al. (2012) and Texas Education Agency (2013) of how to integrate cut-score neighborhoods into a bookmark/item-mapping standard setting.

Feedback Data

If the cut-score neighborhoods have already been established before the standard-setting meeting, it is usually not necessary to share results from all the empirical studies with the standard-setting committee. Results from a few select studies, however, can provide committee members with some context about their cut-score recommendations, help frame their discussion between rounds, and inform their judgments in the rounds to follow.

Because the standard-setting committees are composed primarily of educators and content specialists, they may not be well-versed in statistical concepts such as probability or regression. Thus, for feedback data based on empirical studies to be informative, the data should be presented and explained in a clear, concise and easy-to-understand way. Figure 6 provides an example of feedback data from a STAAR EOC standard-setting committee meeting (Texas Education Agency, 2013).



STAAR English I Reading—Round 1

Performance Standard	Level II	Level III
Borderline Student		
Probability of reaching the corresponding cut in English II Reading	54%	24%
Typical Student		
Probability of reaching the corresponding cut in English II Reading	74%	49%

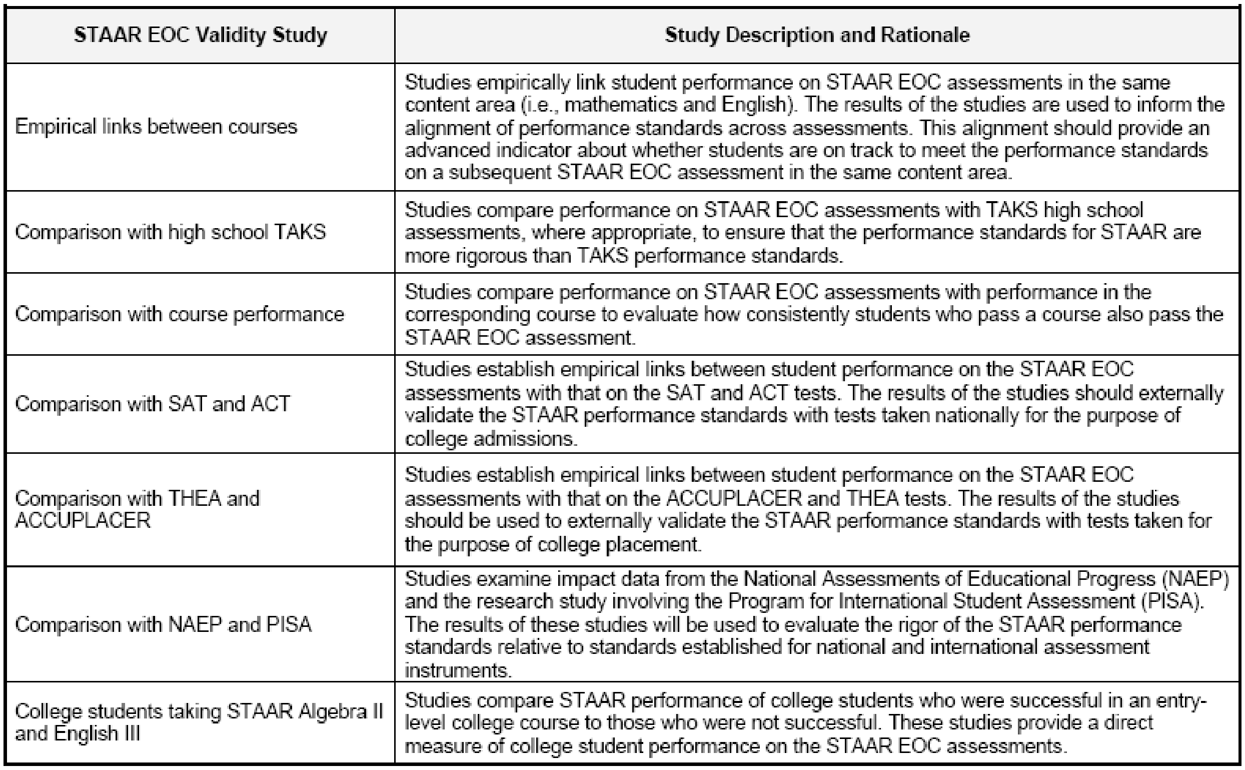
Figure 6. Example Feedback Data for a STAAR EOC Standard-Setting Meeting

In this illustration, the committee had already recommended its cut score for the STAAR English II reading assessment. The feedback data shown was given to the committee after its first round of judgments for STAAR English I reading. It tells the committee how students classified as Level II (Satisfactory) and Level III (Advanced) in English I reading based on the committee's current cut-score recommendation are projected to fare in English II reading in the following year. The feedback data distinguished between borderline students, defined as students who are just barely in the performance category (Level II or Level III), and typical students, defined as students who are in the middle of the performance category. It quantifies the projected success rate of each type of student as probabilities. Committee members can use this feedback information to check the reasonableness of their cut-score recommendations and consider making adjustments, if the adjustments can be supported by their content-based judgments. For example, committee members may consider the 24% probability of a Level III (Advanced) student in English I reading reaching Level III (Advanced) in English II reading as too low. This implies that the current recommended Level III cut score for English I reading may be too low. Consequently, they would consider recommending a more rigorous cut score in the next round, if such an adjustment can be justified by the content they are reviewing.

Lessons Learned

Compared to traditional standard-setting methods, the main element that the EBSS approach introduces to the standard-setting committee is the cut-score neighborhoods. As such, in this subsection, we provide lessons learned from the various EBSS implementations about integrating neighborhoods into the standard-setting meeting.

Because the standard-setting committee members are being asked to work within the neighborhoods, it is important to be transparent about what the neighborhoods are and how they were established. Providing the committee members with a high-level summary of the studies can be useful. Figure 7 gives an example of the empirical studies summary provided to the STAAR EOC standard-setting committee (O'Malley et al., 2012; Texas Education Agency, 2013).



STAAR EOC Validity Study	Study Description and Rationale
Empirical links between courses	Studies empirically link student performance on STAAR EOC assessments in the same content area (i.e., mathematics and English). The results of the studies are used to inform the alignment of performance standards across assessments. This alignment should provide an advanced indicator about whether students are on track to meet the performance standards on a subsequent STAAR EOC assessment in the same content area.
Comparison with high school TAKS	Studies compare performance on STAAR EOC assessments with TAKS high school assessments, where appropriate, to ensure that the performance standards for STAAR are more rigorous than TAKS performance standards.
Comparison with course performance	Studies compare performance on STAAR EOC assessments with performance in the corresponding course to evaluate how consistently students who pass a course also pass the STAAR EOC assessment.
Comparison with SAT and ACT	Studies establish empirical links between student performance on the STAAR EOC assessments with that on the SAT and ACT tests. The results of the studies should externally validate the STAAR performance standards with tests taken nationally for the purpose of college admissions.
Comparison with THEA and ACCUPLACER	Studies establish empirical links between student performance on the STAAR EOC assessments with that on the ACCUPLACER and THEA tests. The results of the studies should be used to externally validate the STAAR performance standards with tests taken for the purpose of college placement.
Comparison with NAEP and PISA	Studies examine impact data from the National Assessments of Educational Progress (NAEP) and the research study involving the Program for International Student Assessment (PISA). The results of these studies will be used to evaluate the rigor of the STAAR performance standards relative to standards established for national and international assessment instruments.
College students taking STAAR Algebra II and English III	Studies compare STAAR performance of college students who were successful in an entry-level college course to those who were not successful. These studies provide a direct measure of college student performance on the STAAR EOC assessments.

Figure 7. Empirical Studies Summary for the STAAR EOC Standard-Setting Committee

In some cases, certain standard-setting committee members may want to make cut-score recommendations outside the established neighborhood boundaries. This can be a delicate situation and each testing program should carefully consider how to strike the balance between maintaining the validity of the EBSS process while giving standard setters the freedom to make their own content-based judgments. For example, the standard-setting facilitator could allow a standard setter to make cut-score recommendations outside of the neighborhood. The standard setter, however, would need to provide a content-based justification on his or her recommendation form as to why this exception is necessary.

Finally, for the sake of continuity, it may be helpful to invite a few policy experts from the group or committee that recommended the neighborhoods to attend the standard-setting committee meetings. These policy experts can help articulate the rationale behind the recommended neighborhoods and serve as advocates of the standard-setting process up to this point. However, it is important to realize that most policy experts are not SMEs and therefore may not be able to appropriately engage in all the tasks required of the standard setters. Thus, if the choice is made to invite policy experts, it may be advisable to include them as observers to the cut-score recommendation process, define specific guidelines about their involvement in the meetings, and clearly communicate the guidelines to each policy expert prior to the meetings.

Summary

This paper discusses the decision points that testing programs face as they implement the EBSS approach to establish performance standards for their assessments. Lessons learned from four implementations of EBSS are synthesized and compared in four key areas: empirical studies, assessed curriculum, policy considerations, and the standard-setting committees. Each implementation of the EBSS approach required slightly different decisions to meet the needs of each testing program. However, some overarching themes exist and can be incorporated into future implementations.

All key areas highlighted in the decision-making process require high levels of planning. All standard-settings require planning, but the incorporation of empirically based studies needs to be laid out far in advance of the actual standard-setting meeting. Decisions around the types of studies to include are dependent on the availability of data for conducting the necessary studies. If data for a study are not available in time to inform the standard setting that study is no longer useful to the standard-setting process. Additionally, if content analyses are used, SMEs need time to plan out which types of analyses will provide the most helpful information. Advance planning of whom to include as policy advisors is vital since people in these types of roles usually have very high demands on their time. It may be necessary to include backup members in case those originally selected to participate are unavailable. The additional steps included in the EBSS approach require careful incorporation into the actual standard-setting committee meeting. This takes additional time due to the additional elements that are included in the process.

In depth knowledge of the testing program for which the standard setting is being conducted is something that is required in all areas of the EBSS approach. It is impossible to make decisions about the types of empirical studies to conduct unless the decision makers have a clear understanding of the purpose of the testing program in question. SMEs need a deep understanding of the assessed curriculum in order to determine which content analyses would be beneficial in helping make decisions about cut score placements. Policy experts bring a unique perspective to the process, but cannot truly provide informed advice without knowing the particulars of the testing program.

And finally, communication is vital at all points throughout the EBSS process. Close discussion with representatives of the testing program must occur to make sure that the types of studies informing the process are truly reflective of the program needs. Representatives of the testing program must be involved in decision making at all points throughout the planning and implementation of the EBSS. Without a clear understanding of the steps of the process, SMEs will not be able to design content-based analyses that contribute to locating appropriate performance standards. If the empirical study information is not clearly explained to both the policy experts and the standard-setting panelists the information will not aid in the decision-making process of either group. Throughout, the importance of using visuals as a way to communicate information has been a key element of EBSS.

By paying careful attention to planning, the needs of the testing program, and communication, the decision making needed to conduct a successful EBSS becomes more obtainable. This evaluation of four key areas of the EBSS process provides valuable insight into what needs to be considered while planning the standard setting. The lessons learned from these different EBSS implementations should provide helpful guidelines and practical advice to the field.

Appendix 1 – Example Study Profile

(*Source: Texas Education Agency, 2013*)

Study Profile: STAAR English III Reading – ACCUPLACER Reading (★★⯨☆☆)

The STAAR English III reading – ACCUPLACER reading external validity study is designed to establish empirical links between performance on the STAAR English III reading test and performance on the ACCUPLACER reading test.

Motivation (★☆☆☆☆)

This analysis was based on a single group of students who took both the STAAR English III reading and the ACCUPLACER reading assessments in 2010 or 2011. Data from STAAR derive from a stand-alone field test administered in 2011 and are linked to a motivated sample of students taking the ACCUPLACER reading test in corresponding years.

Representativeness (★★★★☆) and Sample Size (★★★☆☆)

Grade Levels

All English III Reading Examinees Versus Those Linked to ACCUPLACER Scores

Group	*Grade 8*		*Grade 9*		*Grade 10*		*Grade 11*		*Grade 12*		*Missing*		***Total***
All English	1	0%	85	0%	1,448	4%	33,936	94%	786	2%	7	0%	**36,263**
Linked	0	0%	2	0%	9	1%	1,533	96%	48	3%	0	0%	**1,592**

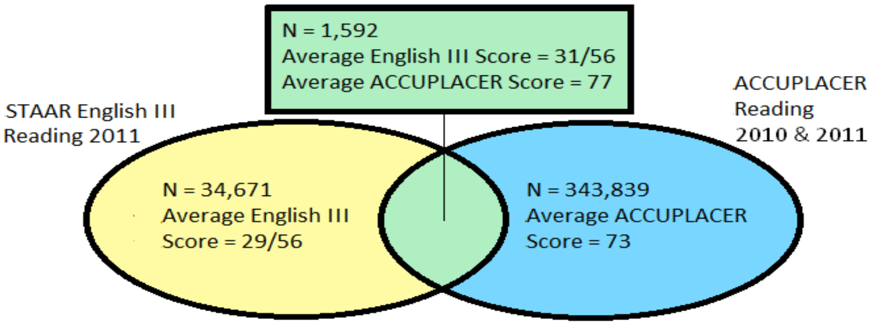
Demographic Characteristics

All English III Reading Examinees Versus Those Linked to ACCUPLACER Scores

Group	*Female*		*Economically Disadvantaged*		*African American*		*Hispanic*		*White*		*Other*	
All English	18,727	52%	16,304	45%	4,815	13%	15,359	42%	13,906	38%	2,183	6%
Linked	911	57%	703	44%	110	7%	767	48%	665	42%	50	3%

Summary of STAAR English III Reading and ACCUPLACER Reading Achievement

Linked and Unlinked Groups



Correlation (★★★☆☆)

Correlation between STAAR English III reading and ACCUPLACER reading = **0.60**

Content Overlap (★★☆☆☆)

There is minimal (approximately 25%) content/skills overlap between the STAAR English III reading assessment and the ACCUPLACER reading test.

Assessment Characteristics

Assessment Characteristic	STAAR English III Reading	ACCUPLACER Reading
Purpose	Created to determine mastery of the English III Texas Essential Knowledge and Skills (TEKS), the state-mandated curriculum.	Under the Texas Success Initiative (TSI), the ACCUPLACER test is used to measure academic skills of students enrolling in Texas institutions of higher education to determine course placement.
Assessment Type	A criterion-referenced assessment	A criterion-referenced assessment
Content	• Measures understanding and analysis of literary, informational, and cross-genre texts • Includes fiction, poetry, drama, literary nonfiction, expository, persuasive, media literacy, and procedural texts • Core skills include using vocabulary in context, making complex inferences and conclusions, analyzing author's craft, and understanding purpose.	• Measures skill level in reading • Short expository passages are used to test reading comprehension • Core skills include identifying details, summarizing, and making inferences and conclusions • Addresses approximately 25% of the content assessed on the STAAR English III assessment, primarily in reading comprehension
Item Format	40 items total: 38 multiple choice items and 2 short answer items	20 multiple choice items total
Administration	• Administered in March, July, and November • Administered by school personnel • Administered online and on paper • Four hour time limit	• Administered on a schedule determined by colleges • Administered by colleges • Administered online: computer-adaptive test • The test is untimed
Performance Standards	Performance standards established and implemented in spring 2012	Under TSI, the scaled passing score for ACCUPLACER reading is 78 (maximum score of 120). Institutions are allowed to set higher standards independent of this passing score.

Appendix 2 – Example Data Quality Summary

(*Source: Texas Education Agency, 2013*)

STAAR EOC Standard Setting Policy Committee: External Validity Studies

Study Name	Motivation	Representativeness	Sample Size	Correlation	Content Overlap	Overall
Algebra II – ACT Mathematics	★★★☆☆	★★★★☆	★★★★★	★★★☆☆	★★★☆☆	★★★⯪☆
Algebra II – SAT Mathematics	★★★☆☆	★★★★☆	★★★★★	★★★☆☆	★★★☆☆	★★★⯪☆
Algebra II – THEA Mathematics	★★★☆☆	★★★★☆	★★★☆☆	★★☆☆☆	★★★☆☆	★★★☆☆
Algebra II – ACCUPLACER Algebra	★★★☆☆	★★★☆☆	★★★★☆	★★☆☆☆	★★★☆☆	★★★☆☆
College Students Take STAAR Algebra II	★★★☆☆	★★★☆☆	★★☆☆☆	★★★☆☆	N/A	★★★☆☆
English III Reading – ACT Reading	★☆☆☆☆	★★★☆☆	★★★★★	★★☆☆☆	★★★☆☆	★★⯪☆☆
English III Reading – SAT Critical Reading	★☆☆☆☆	★★☆☆☆	★★★★★	★★★☆☆	★★★☆☆	★★★☆☆
English III Reading – THEA Reading	★☆☆☆☆	★★★★☆	★★★☆☆	★★☆☆☆	★★★☆☆	★★⯪☆☆
English III Reading – ACCUPLACER Reading	★☆☆☆☆	★★★★☆	★★★☆☆	★★★☆☆	★★★☆☆	★★★☆☆
College Students Take STAAR English III Reading	★☆☆☆☆	★☆☆☆☆	★★☆☆☆	★☆☆☆☆	N/A	★☆☆☆☆
English III Writing – ACT English	★☆☆☆☆	★★☆☆☆	★★★★★	★★★☆☆	★★★☆☆	★★★☆☆
English III Writing – SAT Writing	★☆☆☆☆	★☆☆☆☆	★★★★★	★★★☆☆	★★★☆☆	★★⯪☆☆
English III Writing – THEA Writing	★☆☆☆☆	★★☆☆☆	★★★☆☆	★☆☆☆☆	★★★☆☆	★★☆☆☆
English III Writing – ACCUPLACER Sentence Skills	★☆☆☆☆	★★★★☆	★★★☆☆	★★★☆☆	★★★☆☆	★★★☆☆
English III Writing – ACCUPLACER Written Essay	★☆☆☆☆	★★★★☆	★★★☆☆	★☆☆☆☆	★★★☆☆	★★☆☆☆
College Students Take STAAR English III Writing	★☆☆☆☆	★☆☆☆☆	★★☆☆☆	★☆☆☆☆	N/A	★☆☆☆☆

Study Name	Motivation	Representativeness	Sample Size	Correlation	Content Overlap	Overall
Biology – ACT Science	★★★☆☆	★☆☆☆☆	★★★★★	★★★☆☆	★★★☆☆	★★★☆☆
Biology – SAT Mathematics	★★★☆☆	★☆☆☆☆	★★★★★	★★★☆☆	★★★☆☆	★★★☆☆
Chemistry – ACT Science	★★★☆☆	★★★☆☆	★★★★★	★★★☆☆	★★★☆☆	★★★⯪☆
Chemistry – SAT Mathematics	★★★☆☆	★★☆☆☆	★★★★★	★★★★☆	★★★☆☆	★★★⯪☆
Physics – ACT Science	★★★☆☆	★★★☆☆	★★★★★	★★★☆☆	★★★☆☆	★★★⯪☆
Physics – SAT Mathematics	★★★☆☆	★★★☆☆	★★★★★	★★★☆☆	★★★☆☆	★★★⯪☆
World Geography – ACT Reading	★★★☆☆	★★☆☆☆	★★★★★	★★★☆☆	★★★☆☆	★★★☆☆
World Geography – SAT Critical Reading	★★★☆☆	★☆☆☆☆	★★★★★	★★★★☆	★★★☆☆	★★★⯪☆
U.S. History – ACT Reading	★★★☆☆	★★★☆☆	★★★★★	★★★☆☆	★★★☆☆	★★★⯪☆
U.S. History – SAT Critical Reading	★★★☆☆	★★☆☆☆	★★★★★	★★★★☆	★★★☆☆	★★★⯪☆

Legend	
Motivation	
☆☆☆☆☆	All data (STAAR assessments and external assessments) derive from low-stakes, unmotivated administrations
★☆☆☆☆	Data derive from stand-alone STAAR field tests only, are linked to motivated external assessments, and include constructed responses or highly advanced content
★★☆☆☆	Data derive from stand-alone field tests and low-stakes operational STAAR administrations, are linked to motivated external assessments, and include constructed responses or highly advanced content
★★★☆☆	Data derive from stand-alone field tests and low-stakes operational STAAR administrations, are linked to motivated external assessments, and do not include constructed responses or highly advanced content
★★★★☆	Data derive from some low-stakes and some high-stakes STAAR administrations, are linked to motivated external assessments, and may include constructed responses or highly advanced content
★★★★★	All data (STAAR assessments and external assessments) derive from high-stakes, motivated administrations

Legend

	Representativeness
☆☆☆☆☆	Demographics/student proficiency in the study sample (linked group) and the STAAR test-taking population are distinctly different
★☆☆☆☆	Demographics/student proficiency in the study sample (linked group) and the STAAR test-taking population have minimal similarities
★★☆☆☆	Demographics/student proficiency in the study sample (linked group) and the STAAR test-taking population have some similarities
★★★☆☆	Demographics/student proficiency in the study sample (linked group) and the STAAR test-taking population are moderately similar
★★★★☆	Demographics/student proficiency in the study sample (linked group) and the STAAR test-taking population are very similar
★★★★★	Demographics/student proficiency in the study sample (linked group) and the STAAR test-taking population match perfectly

	Sample Size	*Correlation*	*Content Overlap*
☆☆☆☆☆	0 – 99	0 – 0.39	No relationship
★☆☆☆☆	100 – 499	0.40 – 0.49	Same content area, but no content/skills overlap
★★☆☆☆	500 – 999	0.50 – 0.59	Minimal content/skills overlap (1-25%)
★★★☆☆	1,000 – 1,999	0.60 – 0.69	Some content/skills overlap (26-50%)
★★★★☆	2,000 – 2,999	0.70 – 0.79	Moderate content/skills overlap (51-75%)
★★★★★	3,000 +	0.80 +	Strong content/skills overlap (76-100%)

Appendix 3 – Example Validity Studies Crosswalk

(*Source: O'Malley, Keng, & Miles, 2012*)

Spring 2008 Raw Score	Spring 2009 Raw Score	State Concurrent (Proficiency Levels)	National Concurrent	Predictive Study	Contrasting Groups (Predictive data)	Judgment Studies	Mapping to PLDs
0	0						
…	…						
12	10	ST1-Prof		C or better in CC AL			
13	11	ST2-Prof					
15	12	ST3 & ST4-Prof		C or better in 4S AL			
16	13	ST5-Prof					
17	14						
18	15						
19	16						
20	17	ST6-Prof					
21	18	ST3-Adv		C or better in 4T PC			
23	19	ST1 & ST4-Adv	SAT-Concordance		4-Year Selective	CC AL Cut	
24	20	ST5-Adv	SAT-Concordance			All CC Cut & All AL Cut	
25	21		ACT-Concordance		Community College		
26	22		ACT-Concordance		4-Year Typical	4S AL Cut	
27	23						
28	24						
29	25	ST2-Adv				4T AL Cut	
30	26						
32	27		ACT-Exp. & Pred. Score	B or better in 4S AL & CC PC		All 4S Cut	Prepared
33	28					CC PC Cut & All 4T Cut	
34	29					All PC Cut	
35	30						
36	31		SAT-Exp. & Pred. Score				
37	32		SAT-Pred. Score	B or better in 4T AL			
39	33		SAT-Exp. & Pred. Score	B or better in 4T PC		4T & 4S PC Cut	
40	34	ST6-Advanced	SAT-Pred. Score	A or better in 4S AL			

Spring 2008 Raw Score	Spring 2009 Raw Score	State Concurrent (Proficiency Levels)	National Concurrent	Predictive Study	Contrasting Groups (Predictive data)	Judgment Studies	Mapping to PLDs
41	35	ST6-Advanced	SAT-Exp. & Pred. Score				
42	36	ST6-Advanced	SAT-Exp. & Pred. Score				
43	37	ST6-Advanced					Well Prepared
…	…						
52	45			A or better in 4T PC			
53	46						
54	47						
55	48						
56	49		PSAT-Concordance				
57	50		PSAT-Concordance				
58	51		PSAT-Concordance				
59	52		PSAT-Concordance				
…	…						
76	70						

Legend

- AL = Algebra
- PC = Pre-Calculus
- CC = Community College
- ST1-ST6 = Six States with Proficiency Levels on Crosswalk
- 4T = 4-year Typical Admittance Rate Institution
- 4S = 4-year More Selective Admittance Rate Institution

References

Angoff, W.H. (1971). Scales, norms and equivalent scores. In R.L. Thorndike (Ed.), *Educational Measurement*. Washington DC: American Council on Education.

Beimers, J.N., Way, W.D., McClarty, K.L., & Miles, J.A. (2012). *Evidence based standard setting: Establishing cut scores by integrating research evidence with expert content judgments*. Pearson Bulletin, January 2012, Issue 21. Retrieved from www.pearsonassessments.com.

Boyd, A., Davis, L.L., Powers, S., Schwartz, R., & Phan, H. (2014). *Evidence-based standard setting: vertically aligning grades 3–8 assessments*. Paper presented at the annual meetings of the National Council on Measurement in Education. Philadelphia, PA.

Dorans, N. J., Lyu, C. F., Pommerich, M., & Houston, W. M. (1997). Concordance between ACT Assessment and recentered SAT I sum scores. *College and University, 73*(2), 24-32.

Ferrara, S., Lewis, D., Mercado, R., D'Brot, J., Barth, J., & Egan, K. (2011, April). *A method for setting benchmarked performance standards: Workshop procedures, panelist judgments, and empirical results*. Paper presented at the annual meetings of the National Council on Measurement in Education. New Orleans, LA.

Haertel, E. H., Beimers, J. N., & Miles, J. (2012). The briefing book method. In G. J. Cizek (Ed.), *Setting performance standards* (2nd ed.). New York: Routledge.

Haertel, E. H. (1999). Validity arguments for high-stakes testing: In search of the evidence. *Educational Measurement: Issues and Practice, 18*(4), 5-9.

Iacus, S. M., King, G., & Porro, G. (2011). Multivariate Matching Methods That Are Monotonic Imbalance Bounding. *Journal of the American Statistical Association, 106*(493), 345-361.

Keng, L., Murphy, D., & Gaertner, M. (2012). *Supported by Data: A Comprehensive Approach for Building Empirical Evidence for Standard Setting*. Paper presented at the annual meetings of the National Council on Measurement in Education. Vancouver, BC.

Kingston, N.M., Kahl, S.R., Sweeney, K., & Bay, L. (2001). Setting performance standards using the body of work method. In G.J.Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 219-248). Mahwah, NJ: Lawrence Erlbaum.

LaSalle, A., Munoz, C., Ruff, L., Weisman, E., Sedillo, R., & Phillips, L. (2012). *Grounded in the Content: The Role of Content Analysis in Evidence-Based Standard Setting*. Paper presented at the annual meetings of the National Council on Measurement in Education. Vancouver, BC.

McClarty, K. L., & Davis, L. L. (2012). *Enriched by Policy: Making Performance Standards Meaningful for Educational Outcomes*. Paper presented at the annual meetings of the National Council on Measurement in Education. Vancouver, BC.

Lewis, D. M., Mitzel, H. C., Green, D. R., & Patz, R. J. (1999). *The bookmark standard setting procedure*. Monterey, CA: McGraw-Hill.

O’Malley, K., Keng, L., & Miles, J. (2012). Using validity evidence to set performance standards. In G. J. Cizek (Ed.), *Setting performance standards* (2nd ed.). New York: Routledge.

Phillips, G. W. (2011). The Benchmark Method of standard setting. In G. J. Cizek (Ed.), *Setting performance standards* (2nd ed.). New York: Routledge.

Pommerich, M., Hanson, B. A., Harris, D. J., & Sconing, J. A. (2004). Issues in conducting linkages between distinct tests. *Applied Psychological Measurement, 28*(4), 247-273.

Powers, S., Williams, N., Keng, L., & Starr, L. (2014). *An example of evidence-based standard setting for an English language proficiency test.* Paper presented at the annual meetings of the National Council on Measurement in Education. Philadelphia, PA.

Rosenbaum, P., R., & Rubin, D., B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55.

Texas Education Agency (2013). *State of Texas Assessments of Academic Readiness (STAAR) Assessments Standard Setting Technical Report*. Report published by the Texas Education Agency (TEA) in March 2013. Retrieved from http://www.tea.state.tx.us/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=25769804117&libID=25769804117

Tong, Y., Patterson, B., Swerdzewski, P., Shyer, C. (2014). *Standard Setting for a Common Core Aligned Assessment.* Paper presented at the annual meetings of the National Council on Measurement in Education. Philadelphia, PA.

Webb, N. L., (1997). *Criteria for alignment of expectations and assessments in mathematics and science education*. Washington D.C.: Council of Chief State School Officers and National Institute for Science Education Research Monograph No. 6.

Williams, N., Keng, L., & O'Malley, K. (2012). *Maximizing panel input: incorporating empirical evidence in a way the standard-setting panel will understand*. Paper presented at the annual meetings of the National Council on Measurement in Education. Vancouver, BC.