Evidence-Based Standard Setting: Vertically Aligning Grades 3–8 Assessments

NCME
Philadelphia, PA

Aimee Boyd, Ph.D.
Laurie Laughlin Davis, Ph.D.
Sonya Powers, Ph.D.
Robert Schwartz, Ph.D.
Ha Phan, Ph.D.
April 2014

ALWAYS LEARNING PEARSON

Abstract

Evidence-Based standard setting (EBSS) has been previously used to support alignment of high school assessments to postsecondary expectations. This study presents an extension of the EBSS process to grades 3–8 assessments using the State of Texas Assessments of Academic Readiness (STAAR). This process consisted of empirical studies, policy considerations, and educator committees. Empirical studies included internal links between grades 3–8 and links to the end-of-course (EOC) assessments conducted using matched student data. STAAR grades 3–8 reading and mathematics assessments were also aligned through a vertical scale. External empirical studies indicated the alignment of the STAAR grade 8 assessments and grade 7 writing assessment to national assessments that were aligned to postsecondary readiness. Empirical studies were used before, during, and after the standard-setting meetings to inform the recommended performance standards. The results of the studies were used to establish neighborhoods, or reasonable ranges, for setting performance standards, facilitate the development of ordered item booklets (OIBs), support the committees' recommended performance standards through feedback data, and review the recommended performance standards for reasonableness.

Keywords: evidence-based standard setting, vertical scale, alignment, empirical studies

Evidence-Based Standard Setting: Vertically Aligning Grades 3–8 Assessments

In 2014 the projected enrollment for U.S. public schools from kindergarten through high school is 50.4 million students (National Center for Education Statistics, 2013). The majority of these students will take at least two standardized assessments (reading and mathematics) during the 2013–2014 school year. Each assessment has performance standards that are used to classify students into meaningful categories, including passing status, based on the level of knowledge, skills, and abilities that students demonstrate. Students' achievements on these assessments relative to the performance standards may determine pass rates, graduation rates, and federal and state accountability ratings. Since the passage of No Child Left Behind (No Child Left Behind [NCLB], 2002) the uses and interpretations of performance standards on standardized assessments have varied greatly and grown in intended and unintended consequences. Because of these broad uses, policymakers may have expectations that performance standards should have greater meaning than simply mastery of the subject matter in the current grade. Specifically, policymakers have expressed a desire to have performance standards that indicate students' readiness for the next grade or course (including college courses) or students' preparedness for 21st century skills (including work-force readiness). A change in how performance standards are being interpreted requires standard setting methods that can support these interpretations. Evidence-based standard setting (EBSS) integrates evidence from systematic research with content-based standard-setting procedures to support validity arguments for the interpretation of assessment results (McClarty, Way, Porter, Beimers, & Miles, 2013). The EBSS process consists of empirical studies, policy considerations, and

educator committees which serve as cornerstones in developing, implementing, and evaluating the recommended performance standards.

Texas implemented an evidence-based standard setting (EBSS) approach to establish performance standards for the State of Texas Assessments of Academic Readiness (STAAR) administered at grades 3–8 and high school. The STAAR program is designed to be a comprehensive system, with curriculum and performance standards vertically aligned within high school linking back to middle and elementary school (grades 3–8) and projecting forward to postsecondary readiness. Starting with the capstone end-of-course (EOC) assessments, Algebra II, English III reading and English III writing, performance standards were aligned with postsecondary success through an EBSS process and then extended down to the remaining EOC assessments in order to align performance standards within content areas (Keng, Murphy, & Gaertner, 2012; LaSalle, Munoz, Weisman, Sedillo, & Phillips, 2012; McClarty & Davis, 2012; Texas Education Agency [TEA], 2013; Williams, Keng, & O'Malley, 2012).

The extension of the EBSS process to the STAAR grades 3–8 assessments resulted in vertically aligned performance standards that anchor at high school and link back to middle school then elementary school. The EBSS process for STAAR grade 8 assessments and the grade 7 writing assessment aligned the performance standards for middle school to high school assessments. The STAAR grade 8 assessments and the grade 7 writing assessment then served as anchors to which the grades 3–7 performance standards were linked. This paper discusses the extension of the EOC EBSS process for aligning the STAAR grades 3–8 performance standards including:

- the use of empirical studies to inform the reasonable ranges or "neighborhoods" where performance standards might be recommended;
- the format and type of empirical study results provided to panelists as feedback; and
- the inclusion of study results in a reasonableness review for the recommended performance standards.

State of Texas Assessments of Academic Readiness

The STAAR grades 3–8 assessments were administered for the first time in spring 2012 as a result of the 80th and 81st sessions of the Texas Legislature which called for a new state assessment program to replace the Texas Assessment of Knowledge and Skills (TAKS). Texas students are assessed at grades 3–8 for reading and mathematics, grades 4 and 7 for writing, grades 5 and 8 for science, and grade 8 for social studies. Assessments administered in grades 3–5 are also administered in Spanish. Table 1 lists the content areas assessed in grades 3–8 and the EOC assessments.

Table 1. STAAR Grades 3–8 and EOC Assessments

Grade/Course	Content Area
Grade 3 (English and Spanish)	mathematics and reading
Grade 4 (English and Spanish)	writing, mathematics, and reading
Grade 5 (English and Spanish)	mathematics, reading, and science
Grade 6 (English)	mathematics and reading
Grade 7 (English)	writing, mathematics, and reading
Grade 8 (English)	mathematics, reading, science, and social studies
End-of-Course Assessments	Algebra I, geometry, Algebra II, biology, chemistry, physics, English I reading, English I writing, English II reading, English II writing, English III reading, English III writing, world geography, world history, U.S. history

The STAAR performance standards (Level I: Unsatisfactory Academic Performance, Level II: Satisfactory Academic Performance, and Level III: Advanced Academic Performance) are meant to provide indicators of the degree of preparedness for the next grade level, next course, or postsecondary readiness. These standards provide information not only about what students know and can do but also about their preparedness for future endeavors. When performance standards are set with these goals in mind, it is important to use empirical evidence to validate the interpretations of the standards, including statements about academic content knowledge and the likelihood that students will meet future goals (e.g. success in the next grade level, next course, or postsecondary endeavors). Performance standards for the STAAR 3–8 assessments were set following the first live administration of the tests in spring 2012. Similar to the EBSS process for EOC, the STAAR EBSS process for grades 3–8 consisted of nine steps in order to establish performance standards:

1. Conduct empirical studies
2. Develop performance labels and policy definitions
3. Develop grade/course specific performance level descriptors
4. Develop performance standard ranges
5. Convene standard-setting committees
6. Review performance standards for reasonableness
7. Approve performance standards
8. Implement performance standards
9. Review performance standards

The initial step within the nine-step EBSS process focused on conducting empirical studies. Additional steps incorporated the results of the empirical studies in terms of developing neighborhoods for the performance standards; sharing feedback data with the standard-setting committees; and reviewing performance standards for reasonableness. The EBSS process for STAAR 3–8 was an extension of the EOC EBSS process with adjustments to reflect the vertical alignment of the grades 3–8 assessments, students' readiness for the next grade or course, and to tailor standard-setting committee feedback based on the grades 3–8 empirical studies. The following sections discuss the selection and analysis methods for the empirical studies and the use of the empirical studies throughout the EBSS process.

Empirical Studies Selection and Analysis Methods

Evidence-based standard setting incorporates empirical study results into the standard-setting process to validate the interpretations of student classfications. Potential empirical studies were identified for each STAAR assessment. Selecting the appropriate STAAR 3–8 empirical studies for the EBSS process required reviewing key features for each possible empirical study. Several types of external assessments were considered for inclusion in the empirical studies. These included

- norm-referenced assessments (e.g., Iowa Test of Basic Skills and Stanford Achievement Test),
- assessments linked to college and career readiness (e.g., EXPLORE which links to ACT and ReadiStep which links to SAT), and
- national and international comparative assessments (e.g., NAEP, PISA, TIMSS, and PIRLS)

After some discussion, norm-referenced assessments were removed from consideration as it was unclear to what criterion or benchmark on these assessments the STAAR assessments could be linked. Program for International Student Assessment (PISA), Trends in International Mathematics and Science Study (TIMSS), and Progress in International Reading Literacy Study (PIRLS) were removed from consideration because available data was prior to 2009 and could only be used at the U.S. aggregate level. Of the national and international comparative assessments only National Assessment of Educational Progress (NAEP) had results from 2011 and given that NAEP does not report results at the student level this information could only be used at an aggregate level. EXPLORE and ReadiStep both included a benchmark which links to the college and career readiness benchmark on ACT and SAT, respectively. Also, student level data were available to statistically link the assessments to STAAR.

In addition to studies which linked STAAR to external assessments, internal linkages between STAAR assessments (from grades 7 or 8 to high school and pairwise from grades 3 to 8) were conducted and vertical scales were developed for STAAR 3–8 reading and mathematics as required by legislative statute. Finally, a set of studies comparing student performance on the prior Texas assessment (TAKS) to performance on the STAAR assessment were conducted to allow for policy considerations that the STAAR performance standards be more rigorous than the prior performance standards. Table 2 lists the empirical studies that were selected for use in the STAAR 3–8 EBSS process.

Table 2. Empirical Studies for STAAR Grades 3–8 Assessments

STAAR Assessments	Empirical Studies

STAAR grade 8 mathematics
STAAR grade 8 reading
STAAR grade 7 writing
STAAR grade 8 science
STAAR grade 8 social studies
	• External validity studies
	  o Comparisons with ReadiStep
	  o Comparisons with EXPLORE
	  o Comparisons with NAEP
	• STAAR–TAKS comparison studies
	• STAAR–EOC linking studies

STAAR grades 3–7 mathematics STAAR grades 3–7 reading STAAR Spanish grades 3–5 reading STAAR grade 4 writing STAAR Spanish grade 4 writing STAAR grade 5 science	• External validity studies (comparisons with NAEP) • STAAR–TAKS comparison studies • STAAR–STAAR linking studies
STAAR grades 3–7 mathematics STAAR grades 3–7 reading STAAR Spanish grades 3–5 reading	• STAAR vertical scale studies

STAAR–EOC Linking Study

The STAAR grade 8 assessments and the grade 7 writing assessment were linked to STAAR EOC assessments in order to align performance standards across middle school and high school. For grade 8 reading to English I reading and grade 8 mathematics to Algebra I, the linking was accomplished by first using coarsened exact matching (CEM; Iacus, King, & Porro, 2011) to create randomly equivalent groups. For grade 8 science to biology, chemistry and physics, grade 8 social studies to world geography, world history, and U.S. history and grade 7 writing to English I writing, the linking was accomplished using a single-group design. For more detail on these approaches please see the technical manual (Texas Education Agency, 2013a). Logistic regression was used to compute the probability of attaining a score on the STAAR EOC

assessments given a student's performance on the STAAR grade 8 assessments or grade 7 writing assessment.

Since the STAAR EOC assessments had approved performance standards, the link to EOC assessments was based on the likelihood (60% and 75%) of attaining the Level II performance standard on an EOC assessment given student performance on the STAAR grade 8 assessments and the grade 7 writing assessment. The Level II performance standard states that "students in this category have a reasonable likelihood of success in the next grade or course but may need short-term, targeted academic intervention." A *reasonable likelihood* of success for students in the Level II performance standard on the grade 8 assessments or the grade 7 writing assessment was operationalized as a 60 percent likelihood and *success* was operationalized as attaining Level II in the next grade or course. The Level III performance standard states that "students in this category have a high likelihood of success in the next grade or course with little or no academic intervention." A *high likelihood* of success for students in the Level III performance standard was operationalized as a 75 percent likelihood and *success* was operationalized as attaining Level II in the next grade or course. The Level II performance standard represents the passing standard on STAAR assessments. The likelihood results informed the neighborhoods in which the standard-setting committees set cut scores for STAAR grade 8 assessments and grade 7 writing.

Table 3 lists the data collection design, sample size, and correlation for each of the STAAR-EOC linking studies. Chemistry, physics, world history, and U.S. history are not generally taken by Texas students in grade 9 which limited the sample size for these studies. These students were likely to be a more able student group taking advanced classes compared to the

typical science and social studies assessments (biology and world geography, respectively) taken by Texas students in grade 9.

Table 3. Sample Size and Correlation for STAAR 3–8 Empirical Links to STAAR EOC

STAAR Assessment	Linked Test	Data Collection	Sample Size	Correlation
Grade 8 mathematics	Algebra I	Coarsened exact matching	466,202	0.70
Grade 8 reading	English I reading	Coarsened exact matching	559,998	0.75
Grade 8 reading	English I writing	Coarsened exact matching	559,853	0.77
Grade 8 science	Biology	Single-group design	285,220	0.74
Grade 8 science	Chemistry	Single-group design	1,196	0.73
Grade 8 science	Physics	Single-group design	1,031	0.77
Grade 8 social studies	World geography	Single-group design	286,239	0.77
Grade 8 social studies	World history	Single-group design	4,893	0.74
Grade 8 social studies	U.S. history	Single-group design	1,579	0.78
Grade 7 writing	English I writing	Single-group design	288,511	0.73

STAAR–STAAR Linking Study

Studies empirically linked student performance across grades within content areas for the STAAR 3–8 assessments in order to align performance standards across elementary and middle school grades. Because spring 2012 was the first administration of the STAAR 3–8 assessments, a cohort of students that had taken both STAAR assessments was not available for linking. For reading and mathematics, where assessments were offered at every grade level, links across grades were established using coarsened exact matching where students at the higher grade were matched to students at the lower grade based on a common prior test score (TAKS) in the same content area. For example, to link STAAR grade 4 mathematics to STAAR grade 5 mathematics, both cohorts would be matched using TAKS grade 3 mathematics scores (which is one year removed for the fourth grade cohort and two years removed for the fifth

grade cohort). In these cases, the links were based on large sample sizes and the matching variables indicated strong relationships to STAAR test scores.

For science and writing, where assessments were not offered at every grade level, links across grades were established using coarsened exact matching where students at the higher grade were matched to students at the lower grade based on a common prior test score (TAKS) in reading and mathematics.

For the linking studies from grade 3 to grade 4 reading and mathematics no common prior assessment was available. For example, grade 3 is the first grade tested in the STAAR program; therefore, no common prior assessment was available to link STAAR grade 3 reading to STAAR grade 4 reading through coarsened exact matching. For these analyses, a single-group design based on the STAAR-to-TAKS comparisons was implemented. The STAAR grade 4 reading and mathematics data in 2012 were linked to the same students' TAKS grade 3 reading and mathematics data in 2011. The established link between the STAAR grade 3 and TAKS grade 3 assessments were leveraged for these analyses.

Logistic regression was used to compute the probability of attaining a score on the upper-grade-level STAAR assessment given a student's performance on the lower-grade-level STAAR assessment. These results informed the neighborhoods in which the standard-setting committees recommended performance standards.

STAAR to TAKS Comparison Study

Studies compared performance on STAAR to performance on TAKS in order to evaluate whether the performance standards for STAAR were more rigorous than the TAKS performance

standards. For reading, mathematics, science, and social studies, items for the new STAAR assessment were embedded into field-test positions on the TAKS assessments in 2011. This allowed for a common-item non-equivalent groups design to be used to place the TAKS and STAAR items onto the same scale using the Rasch model (Rasch, 1980). For grades 4 and 7 writing assessments, a separate STAAR stand-alone field test was administered in 2011 rather than embedding STAAR items in the TAKS writing assessments. Since there were no common items and the STAAR writing assessment had extensive changes in the writing items compared to TAKS writing items, equipercentile linking based on a single-group design was implemented. Students who took TAKS writing assessments in spring 2011 and the STAAR stand-alone field test in spring 2011 were included in the sample. For the STAAR-to-TAKS comparison studies in grades 3–8, the TAKS Met Standard performance standards were identified on the STAAR assessments. The empirical results were then evaluated with respect to trends in TAKS impact data and the impact data for the STAAR 2012 assessments. Neighborhoods for the STAAR passing standard were set above the point identified on the STAAR scale as comparable to the TAKS passing standard.

STAAR to External Linking Study

The STAAR grade 8 assessments and the grade 7 writing assessment were linked to external measures—EXPLORE and ReadiStep—which are linked to ACT and SAT, respectively and, therefore, serve as early indicators of postsecondary readiness. Both assessments are typically administered to students in grade 8. Linking was done in a pairwise chained fashion such that EXPLORE and ReadiStep were first linked to TAKS (using a single group design and

data from 2010 and 2011). The link between STAAR and TAKS on the Rasch scale was used to provide the STAAR to External linking study results on the STAAR scale.

Logistic regression was used to compute the probability of attaining a reference point on the EXPLORE and ReadiStep assessments given a student's performance on the STAAR assessments. The reference points for EXPLORE and ReadiStep were established based on the linking relationships in place for the ACT and SAT, respectively. ACT and The College Board linked EXPLORE and ReadiStep to the point on ACT and SAT that represented a 75 percent chance of earning a C or better in a corresponding college course. The points along the STAAR scales indicating that students would be at least 50 percent likely to meet or exceed the reference points for EXPLORE and ReadiStep were then determined and used to inform the neighborhoods in which the standard-setting committee's recommended performance cut scores.

Table 4 lists the sample size and correlation for each of the STAAR to External linking studies. Table 5 lists the benchmarks on the EXPLORE and ReadiStep assessments used in the linking studies.

Table 4. Sample Size and Correlation for STAAR 3–8 External Validity Studies

STAAR Assessment	Linked Test	Sample Size	Correlation
Grade 8 mathematics	EXPLORE mathematics	113,244	0.67
Grade 8 mathematics	ReadiStep mathematics	186,729	0.66
Grade 8 reading	EXPLORE reading	113,240	0.60
Grade 8 reading	ReadiStep critical reading	187,209	0.59
Grade 8 science	EXPLORE science	112,264	0.61
Grade 8 social studies	EXPLORE reading	112,397	0.61
Grade 7 writing	EXPLORE English	108,644	0.66
Grade 7 writing	ReadiStep writing	179,933	0.63

Table 5. Measures and Benchmarks Linked to STAAR 3–8 Assessments

STAAR Assessment	Linked Test	Reference Point
Grade 8 mathematics	EXPLORE mathematics	17
Grade 8 mathematics	ReadiStep mathematics*	3.9
Grade 8 reading	EXPLORE reading	15
Grade 8 reading	ReadiStep critical reading*	2.5
Grade 8 science	EXPLORE science	20
Grade 8 social studies	EXPLORE reading	15
Grade 7 writing	EXPLORE English	13
Grade 7 writing	ReadiStep writing*	1.0

*The reference points for ReadiStep published by the College Board are linked to SAT scores based on first year college GPA. The reference points for ReadiStep in these analyses are linked to SAT scores representing a 75% probability of attaining a grade of C or higher in a corresponding college course.

STAAR Vertical Scale Study

STAAR 3–8 reading and mathematics assessments were placed on a vertical scale, which puts all items and student proficiency on a common scale within a content area across grade levels. The vertical scale allows the comparison of student performance across grades within a content area and was used to inform the alignment of standards for STAAR 3–8 assessments in reading and mathematics. The vertical scale study determined the difference in test difficulty between adjacent grades for STAAR 3–8 reading and mathematics assessments. The data were collected in spring 2012 during the first operational administration of STAAR. The data-collection design was a common item non-equivalent groups design in which students in adjacent grade levels responded to the same items, thereby allowing direct comparison of item difficulties. These vertical linking items were embedded in field-test positions. Both upper-grade-level and lower-grade-level items were included in the design (e.g., grade 4 mathematics items were included in grade 5 mathematics and grade 3 mathematics test forms). The Rasch

model was used to place all items and student proficiencies on the same scale within a content area. Additional information on the STAAR vertical is available in the technical report (Texas Education Agency, 2013b).

Use of Empirical Studies

Empirical studies play a significant role in all steps of the EBSS process. Empirical studies were used before, during, and after the standard-setting meetings to inform the recommended performance standards. The results of the studies were used to establish neighborhoods for setting performance standards, facilitate the development of ordered item booklets (OIBs), support the committees' recommended performance standards through feedback data, and review the recommended performance standards for reasonableness. This section discusses the use of the empirical studies in multiple steps of the EBSS process.

Neighborhood Development

Neighborhoods (which indicate reasonable ranges in which the standard setting committees would recommend performance standards) were established for two cut scores which determined three performance levels (Level I: Unsatisfactory Academic Performance, Level II: Satisfactory Academic Performance, and Level III: Advanced Academic Performance). The STAAR 3–8 neighborhoods were developed based on a set of guidelines which conceptually mapped the different empirical studies (prior to having study results) to one of the three performance levels. For example, the college readiness benchmark for EXPLORE was expected to fall within the neighborhood for Level III: Advanced Academic Performance. Similarly the TAKS passing standard was expected to fall within the neighborhood for Level I: Unsatisfactory

Academic Performance. Once study results were available, neighborhood boundaries could then be drawn which best matched the conceptual expectations.

A horizontal number line, representing students' performance from 100 percent to zero percent meeting the standard, was used to illustrate the neighborhood guidelines and the empirical neighborhood number lines for each STAAR 3–8 assessment for the state. The standard-setting committees were not provided the number lines during standard setting in order to reduce the burden of feedback data. Figure 1 illustrates the STAAR 3–8 neighborhood development guidelines on a horizontal number line. In addition to the empirical studies, general guiding principles were established for neighborhood development. The general guiding principles included:

- performance standards that are aligned with the EOC content areas
- performance standards informed by validity study results
- measurement precision where the cut scores are set
- reasonable raw score cuts
- reasonable impact data

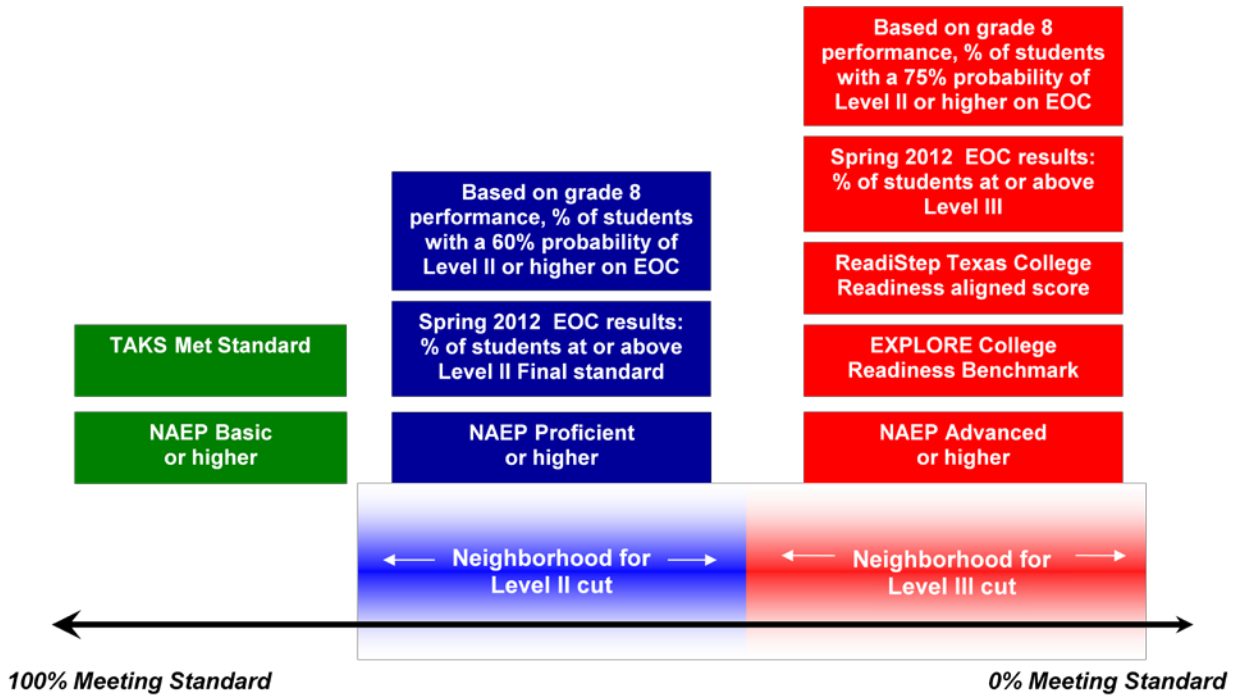

Figure 1. Graphical Illustration of STAAR 3–8 Neighborhood Development Guidelines

Using the empirical study results and the neighborhood development guidelines, an empirical number line was constructed for each assessment. The values displayed on the number line were the percentage of students (based on performance on the spring 2012 administration) who scored at or above this point on the STAAR assessment of interest. This scale metric was chosen so that the percentage of students that would meet or exceed a cut score if it were strictly aligned with the result of a particular study was easily seen. The neighborhoods established for the STAAR grade 8 assessments (reading, mathematics, science, and social studies) and grade 7 writing were used to inform the neighborhoods for the remaining STAAR grades 3–7 assessments by working backward from grade 8 to grade 3. Figure 2 illustrates an example empirical number line for STAAR grade 8 mathematics.

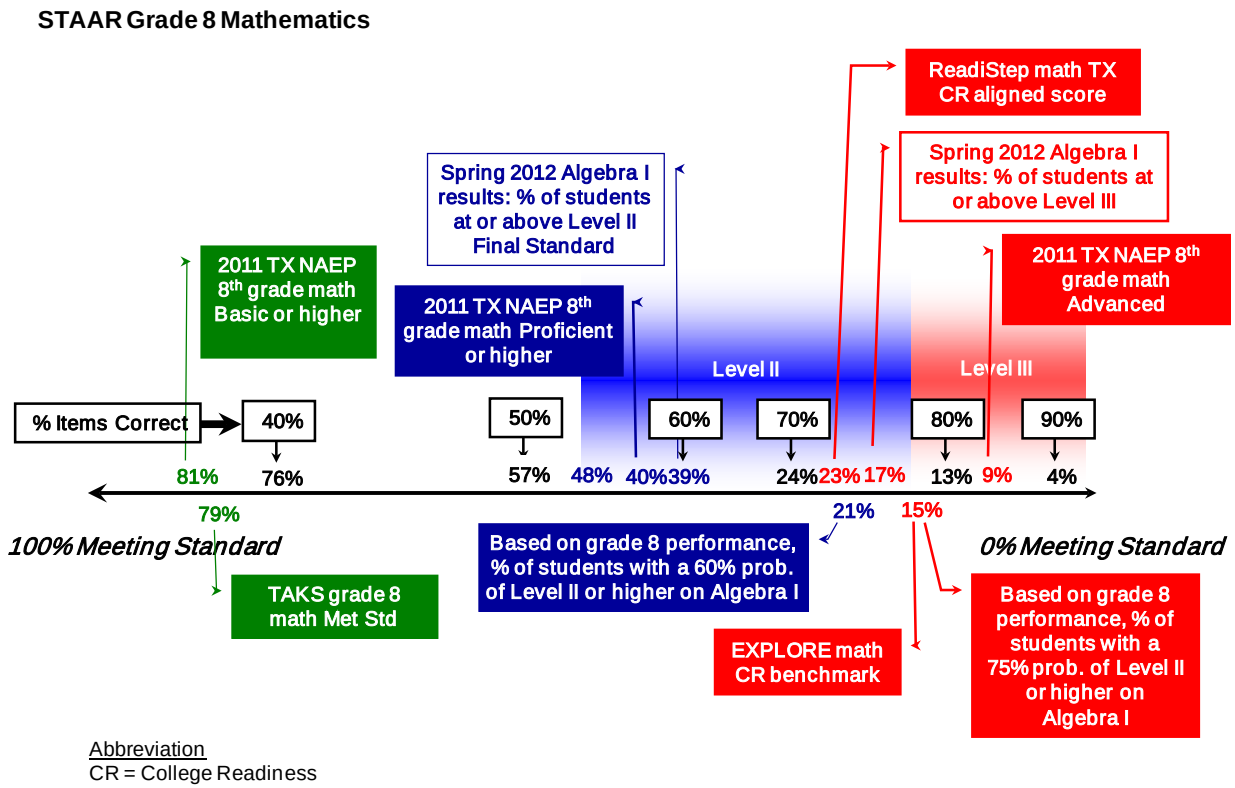

Figure 2. Empirical Number Line for STAAR Grade 8 Mathematics

As can be seen in the figure, not all of the study results aligned cleanly with the expected conceptual mapping. For example, there are some "red boxes" (like the ReadiStep college readiness benchmark) which were study results expected to fall within the Level III neighborhood, but which fell within the Level II neighborhood once neighborhood boundaries were drawn. This could have been addressed by shifting down the lower boundary of the Level III neighborhood to capture all of the red boxes. However, doing so would have created the opposite issue—in this case the "blue box" which indicated the linking study result between grade 8 mathematics and Algebra I would then have fallen into the Level III neighborhood. This

type of mismatch between conceptual expectations and actual study results was also encountered in the development of neighborhoods for STAAR EOC. Making determinations of how to navigate these real data challenges requires careful consideration and weighing of factors such as the data quality from each study, the degree of content overlap between assessments being compared, and the meaning of reference points on the various assessments.

The neighborhoods for STAAR reading and mathematics for grades 3–7 were informed using the alignment of the vertical scale across grades, the STAAR–STAAR linking studies and the STAAR–TAKS comparison studies. Since the assessments within reading and mathematics were on the same vertical scale, the neighborhoods could be graphically displayed using the vertical scale and evaluated to show that the neighborhoods increased as the grades increased. The vertical-scale neighborhoods, such as the one shown in Figure 3, were generated for grades 3–8 for reading and mathematics. The figure shows the lower bound for the Level II and Level III neighborhoods. The impact data (percentage of students meeting or exceeding the Level II and Level III standards) shown in the vertical scale graphic are based on student performance during the spring 2012 administration.

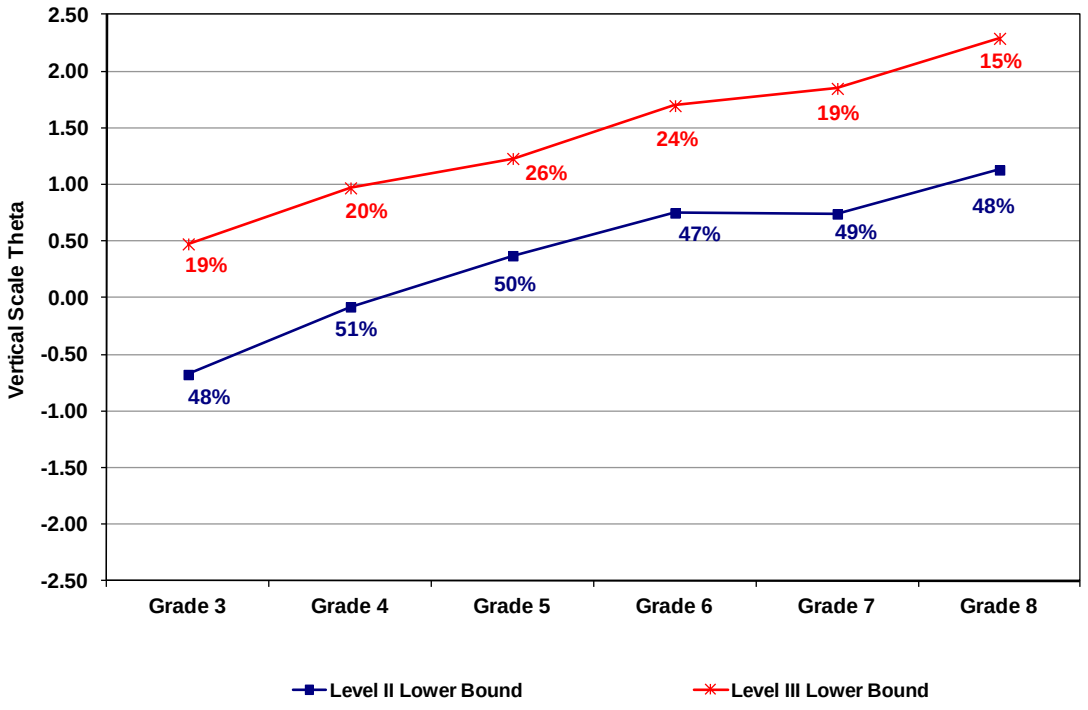

Figure 3. Vertical Scale Graphic for STAAR Grade 3–8 Mathematics Neighborhoods

For STAAR grade 4 writing and grade 5 science, a vertical scale was not available. The neighborhood number lines were generated for STAAR English grade 4 writing, STAAR Spanish grade 4 writing, and grade 5 science assessments based on the upper-grade-level assessment in the same content area. The STAAR-STAAR to linking studies and percent of students in the neighborhood boundaries for the upper-grade-level assessment informed the neighborhoods for these assessments. After the neighborhoods were identified for all STAAR 3–8 assessments, only the neighborhood boundaries were carried forward in the EBSS process in order to reduce the data complexity for the standard-setting committees.

Spanish Language Versions of STAAR

Spanish versions of the STAAR assessments are available in grades 3–5. For grades 3–5 mathematics and grade 5 science, the assessments are transadapted, that is, translated from English into Spanish and adapted as necessary to account for cultural and linguistic differences. As such, the English and Spanish mathematics and science tests share a common scale and a single set of performance standards applies to both language versions of the tests. Assessments administered in grades 3–5 for reading and grade 4 writing are uniquely developed in Spanish. As such, English and Spanish tests are on different scales and require performance standards to be set separately. The performance standards for English and Spanish language versions of STAAR reading and writing assessments were set so as to produce comparable score interpretations.

Linking studies were conducted for the STAAR Spanish grades 3–5 reading assessments and STAAR Spanish grade 4 writing assessment (e.g., Spanish grade 4 writing to STAAR English grade 7 writing) to inform the Spanish neighborhoods. The linking results were less compelling compared to the STAAR English grades 3–5 reading and STAAR English grade 4 writing studies. Impact data is expected to be somewhat lower for the Spanish assessments compared to English assessments because the Spanish population contains a greater proportion of economically disadvantaged students—a group that typically has lower achievement. The Spanish impact data suggested by the linking studies was much lower than the English impact data. Based on the linking studies, the impact data for the STAAR Spanish assessments were much lower, even considering the larger proportion of economically disadvantaged students, compared to the STAAR English assessments. For students testing in Spanish for all

assessments, the expectation was similar impact data across the Spanish mathematics, reading, science, and writing assessments. When the STAAR Spanish reading and writing impact data was compared to the STAAR Spanish mathematics and science impact data (which used the same study results as the STAAR English assessments since the studies were on a common scale) the impact data was much lower reading and writing. Further examination indicated that the nature of the Spanish population changes in terms of size and academic ability. As Spanish-speaking students develop academic proficiency in English, they move from testing in Spanish to testing in English. The Spanish students at grade 5 are systematically different from those at grade 3. These changes may have resulted in inadequate linking results. Therefore, the development of the Spanish neighborhoods focused more on the strength of the test-construction process to guide neighborhood development. The English and Spanish assessments are developed to assess the same grade-level student expectations. The test-development process is designed to result in language versions that are comparable in terms of the content that is measured. The relationship between the two language versions would suggest that a comparable standard would require students to correctly answer a similar number of items. Therefore the linking study results from the English assessments were used to help inform the neighborhoods in which the standard-setting committee's recommended STAAR Spanish performance cut scores.

Ordered Item Booklet

Once neighborhoods were developed for the cut scores, ordered item booklets (OIBs) were created based on the neighborhoods. OIBs consisted of items ordered by Rasch item difficulty, easiest to hardest. Standard-setting panelists reviewed the items in the OIBs and

placed a bookmark following the items that they determined best represented the minimum expected performance for each performance level. Given the importance of the OIB to the standard-setting process, each booklet was carefully constructed to give panelists the most information about the types of items falling within the neighborhoods.

For STAAR assessments, the OIBs were ordered from easiest to hardest based on the Rasch item difficulty values. A sample test form was used as the starting point for each OIB. Since the neighborhoods represented the reasonable range within which the cut scores should fall, items not part of the original test blueprint were added to the OIB in order to increase the number of items within the neighborhood bounds. This allowed panelists to make finer distinctions between items within the neighborhoods.

Areas of the OIB that did not have item representation along the Rasch scale were identified as gaps. Areas of the OIB with an overrepresentation of items along the scale were identified as clusters. This information, as well as the item's Rasch item difficulty, was used to select additional items to fill in gaps in the OIB. For clusters, the number of items appearing in that section of the OIB was reduced.

Standard-Setting Committees

For the STAAR 3–8 assessments, sixteen standard-setting committees recommended performance standards for 21 assessments. The organization of the standard-setting committees into four consecutive meetings allowed the recommended performance standards from the upper grade-level committees to be used as feedback for the lower grade-level committees. In addition, several committees recommended performance standards for more

than one assessment. For these meetings, the committees started with the higher grade level then recommended the lower grade level performance standards. Table 6 lists the organization of the standard-setting committees across the four consecutive meetings.

Table 6. STAAR 3–8 Standard-Setting Committee Meeting Organization

Standard-Setting Meetings	Committees by Subjects and Grades
Meeting Set 1	Mathematics grade 8 Reading grade 8 Writing grade 7 Science grade 8 Social studies grade 8
Meeting Set 2	Mathematics grades 6 and 7 Reading grades 6 and 7
Meeting Set 3	Mathematics grade 5 English reading grade 5 Spanish reading grade 5 Science grade 5
Meeting Set 4	Mathematics grades 3 and 4 English reading grades 3 and 4 Spanish reading grades 3 and 4 English writing grade 4 Spanish writing grade 4

For the STAAR 3–8 standard-setting meetings, each committee had the opportunity to review their recommendations in conjunction with the recommendations of committees that met previously (i.e., higher grade levels). For the grades 3–5 reading and grade 4 writing assessments, the English and Spanish committees met together for the specific performance level descriptors (PLDs) discussion, the borderline student discussion, and the standard-setting training. The committees separated before judgments and feedback occurred.

For each STAAR 3–8 standard-setting meeting, the panelists engaged in three rounds of judgments for each assessment. Within each round, panelists were asked to carefully review the OIB item by item and then make a recommendation for the Level II: Satisfactory Academic Performance cut score first, followed by a recommendation for the Level III: Advanced Academic Performance cut score. During the first round of judgments, the OIB was marked at the lower bound for the Level II neighborhood, although panelists were able to select an item below the lower bound if the content supported the judgment. After making the Level II judgments, the panelists were provided with the lower bound for the Level III neighborhood. Revealing the neighborhood boundaries sequentially allowed panelists to focus on the content of the OIB past each marker without feeling restricted about how far into the OIB they could place their recommendations. After the first round of judgments, panelists could see both the Level II and Level III lower neighborhood boundaries for their Round 2 and Round 3 judgments.

Between the judgment rounds, the panelists were provided information—including empirical study results and impact data— that they used to refine their judgments. After each round of judgment, the following feedback data were presented to the panelists.

Round 1 Feedback Data

- The panelist's individual Round 1 cut-score recommendations (bookmarked pages) for Level II and Level III
- Table-level Round 1 cut-score recommendations—the minimum, maximum, mean, and median bookmarked pages for Level II and Level III
- Committee-level Round 1 cut-score recommendations—the minimum, maximum, mean, and median bookmarked pages for Level II and Level III

- Panelist agreement chart for each committee member's Round 1 cut-score recommendation
- The percentage of students answering each item in the OIB correctly (p-values)

For grade 8 assessments and the grade 7 writing assessment, additional feedback included data showing projections from the committee-level Round 1 cut-score recommendations to the next course in a course-taking sequence or to external measures linked to postsecondary readiness. Figures 4 and 5 are examples of feedback data after a judgment round for grade 8 reading. Information that related scores on the STAAR assessments with other assessments was presented in two ways: in relation to the borderline student (the student just barely making it into a performance level) and in relation to the typical student (the student in the middle of a performance level). The borderline student and typical student were defined by where the cut scores fell after each judgment round. Panelists were asked to consider the reasonableness of the feedback data given the expectation of students in the Level II and Level III performance standards. For example, do the probabilities of reaching Level II and Level III in English I reading for the borderline and typical student reflect the expectations of students as they transition from the grade 8 reading curriculum to English I curriculum.

Performance Standard	**Level II**	**Level III**
Borderline Student		
Probability of reaching the Level II cut score in English I Reading	63	89
Typical Student		
Probability of reaching the Level II cut score in English I Reading	79	95

Figure 4. Example Feedback Data for Grade 8 Reading Projections to the Next Course

Performance Standard	Level II	Level III
	Borderline Student	
Probability of reaching the EXPLORE benchmark	33	49
	Typical Student	
Probability of reaching the EXPLORE benchmark	40	57

Figure 5. Example Feedback Data for Grade 8 Reading Projections to the External Measures

For STAAR grades 3–7 reading and mathematics assessments, additional feedback included vertical scale data showing the committee-level Round 1 cut-score recommendations and the higher grade-level Level II and Level III cut-score recommendations from prior committees. Figure 6 illustrates an example of feedback data including vertical scale results after a judgment round for STAAR grade 5 mathematics. Panelists were asked to consider the progression of the vertical scale cut scores across grades and the progression of the mathematics curriculum across grade levels. For example, the distance between the Level II and Level III cut scores given the performance level descriptors for the borderline students.

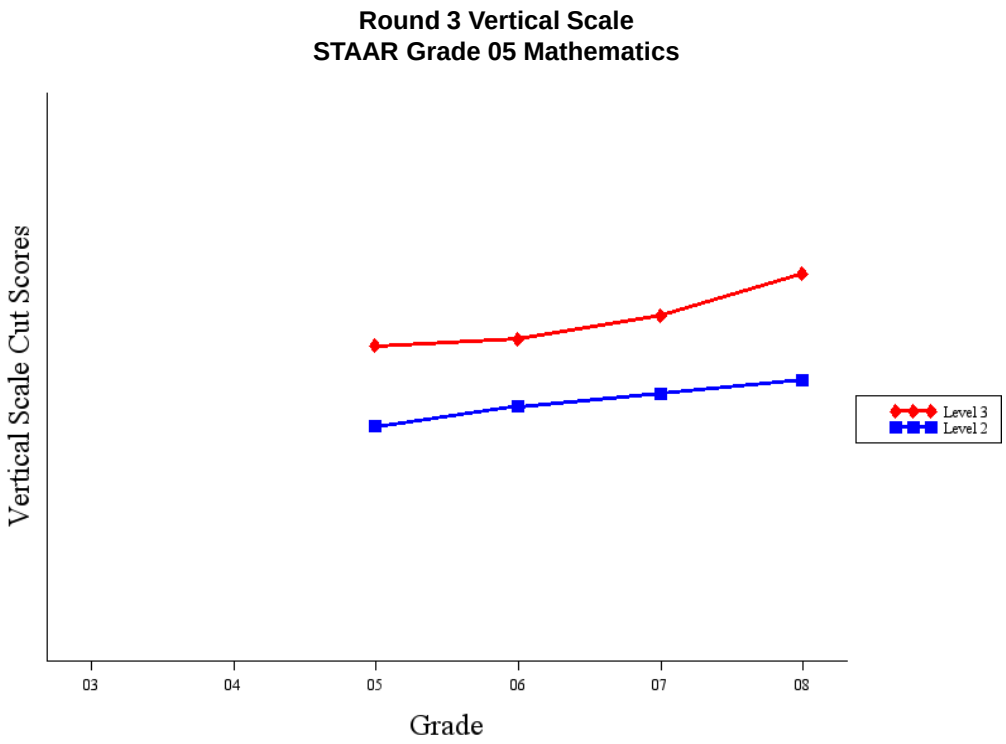

Figure 6. Example Vertical Scale Feedback Data for Grade 5 Mathematics

Even though the grade 4 writing and grade 5 science assessments have higher grade-level assessments in the same content area, they are not vertically scaled, so those committees were not provided this type of feedback. The committees were provided impact data based on the recommended performance standards from the upper-grade-level standard-setting committees. Panelists were asked to consider the impact data for their assessment in comparison to the upper-grade-level impact data in terms of a "reality check" given the differences in the curriculum across grades.

Round 2 Feedback Data

- Round 2 feedback data similar to the data provided after Round 1
- Impact data based on the committee's Round 2 cut-score recommendations

Round 3 Feedback Data

- Round 3 feedback data similar to the data provided after Round 1
- Impact data based on the committee's Round 3 cut-score recommendations

Table 7 lists the types of empirical study feedback presented to the standard-setting committees after all three rounds of judgments.

Table 7. STAAR 3–8 Standard-Setting Committee Empirical Study Feedback

Assessment	Next EOC Course	ReadiStep	EXPLORE	Vertical Scale
Grade 8 Reading, Grade 8 Science, and Grade 7 Writing	**Yes**	*No**	**Yes**	*No*
Grade 8 Mathematics	**Yes**	**Yes**	**Yes**	*No*
Grade 8 Social Studies	**Yes**	*No*	*No*	*No*
Grades 3–7 Reading and Mathematics	*No*	*No*	*No*	**Yes**
Grade 4 writing and Grade 5 science**	*No*	*No*	*No*	*No*

* for grade 8 reading and grade 7 writing committees, the results of the ReadiStep linking studies were not presented because they were outside of the neighborhoods (likelihood would always have been 100% regardless of where the committee made their judgments)
**for grade 4 writing and grade 5 science the committees were provided results based on the recommended performance standards for the grade 7 writing and grade 8 science, respectively

All feedback given to the panelists expressed the cut scores in terms of a page number in the OIB. Panelists were not provided with the raw-score or percent-correct values for a test form associated with their bookmark placement.

Panelist Evaluation Survey

At the end of the standard-setting meeting, panelists were asked to complete a process-evaluation survey. The purpose of the survey was to collect information about each panelist's experience in recommending cut scores for the STAAR assessments. Although there was some variation across STAAR 3–8 committees, most panelists thought that the various components of the meeting were "successful" or "very successful." The majority of panelists thought that the activities conducted during the meeting were either "useful" or "very useful." They also reported that the time spent on training, table discussions, and judgment tasks was "adequate" to "more than adequate." When asked about their confidence in the PLDs and the cut scores, most panelists felt "confident." Virtually all committee members thought that they were given adequate opportunity to express their opinions, ask questions, and interact with other committee members. Additionally, the majority of panelists indicated that they believed that their opinions and judgments were respected by others.

Reasonableness Review

After standard-setting committees recommended Level II and Level III performance standards for the STAAR 3–8 assessments, the state conducted reasonableness reviews of the cut-score recommendations across content areas and made adjustments as appropriate. This process confirmed that the performance standards contributed to a well-articulated and coherent assessment system. For the STAAR 3–8 assessments, the state conducted a reasonableness review of the cut-score recommendations not only within and across 3–8 content areas but also in relation to the STAAR EOC cut scores. The state evaluated the results

from all the standard-setting committees by considering many pieces of information, including the empirical study results associated with Round 3 judgments and the group discussion recommendations.

Summary

The extension of EBSS to the STAAR grades 3–8 assessments resulted in a comprehensive assessment system with vertically aligned performance standards that anchored at high school and linked back to middle school then elementary school. The empirical studies provided evidence to support the standard-setting committees recommended performance standards. The empirical studies included both internal assessments to the Texas program and external assessments. This paper detailed many of the limitations and challenges in using empirical studies, as well as their usefulness in informing the standard-setting process.

The STAAR 3–8 reading and mathematics vertical scale provided a framework for aligning the neighborhoods, developing the OIBs, providing feedback to standard-setting committees in meaningful ways, and reviewing the recommended performance standards for reasonableness. The STAAR grades 4 and 7 writing assessments, STAAR grades 5 and 8 science assessments, and the STAAR grade 8 social studies assessment lacked an underlying vertical scale framework. This limited the empirical studies to the STAAR-STAAR linking studies which also included challenges in terms of data quality and availability given the need to establish performance standards after the first administration.

Limitations in the availability of student-level data for external assessments limited the role of external empirical studies to the STAAR grade 8 assessments and grade 7 writing

assessment. Additional data collection was not feasible for external assessments administered at lower grade levels. Additionally, there were limitations in the internal and external data available for the STAAR grades 3–5 Spanish assessments. The data for the Spanish assessments were not as stable as English data due to the changing populations moving from the Spanish assessments to the English assessments.

Longitudinal data is preferable for reviewing and evaluating the recommended performance standards, but the timing of the standard-setting meetings after the first administration of STAAR assessments limited some of the empirical studies in terms of data quality. This was especially prevalent for the assessments that are not administered in consecutive grade levels.

The implementation of EBSS to the STAAR 3–8 standard-setting process allowed for the integration of evidence from systematic research with content-based standard-setting procedures to support the interpretation of STAAR assessment results. The organization of the standard-setting meetings allowed for committee recommended performance standards for upper grade levels to be used as feedback to the lower grade levels. This structure enhanced the vertical alignment of the performance standards. The strength of the EBSS process comes from the ability to integrate subjective content expert judgment with more objective empirical evidence. This was evident in the panelists' evaluation survey results. The feedback data provided panelists' confidence that the content-based judgments were aligned with the empirical data. However, the quality of the empirical evidence needs to be carefully considered in determining how to weigh these two factors. The quality and availability of data to conduct empirical studies may be less than ideal, especially when an assessment program is new,

instruction of the content is not yet optimized, and longitudinal performance data are not yet available.

References

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, April). *Supported by data: A comprehensive approach for building empirical evidence for standard setting.* Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, British Columbia, Canada.

LaSalle, A., Munoz, C., Weisman, E., Sedillo, R., & Phillips, L. (2012, April). *Grounded in the content: The role of content in evidence-based standard setting.* Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, British Columbia, Canada.

McClarty, K. L., & Davis, L. L. (2012, April). *Enriched by policy: Making performance standards meaningful for educational outcomes.* Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, British Columbia, Canada.

McClarty, K. L., Way, W. D., Porter, A. C., Beimers, J. N., & Miles, J. A. (2013). Evidence-based standard setting: Establishing a validity framework for cut scores. *Educational Researcher, 42(2),* 78-88. doi: 10.3102/0013189X12470855.

National Center for Education Statistics (2013). *Projections of Education Statistics to 2021*. Retrieved from http://nces.ed.gov/pubs2013/2013008.pdf

No Child Left Behind (NCLB) Act of 2001 (2002). Pub. L. No. 107- 110, § 115, Stat. 1425.

Rasch, G. (1980). *Probabilistic Models for Intelligence and Attainment Tests*. Chicago: The University of Chicago Press.

Texas Education Agency (2013a). *State of Texas Assessments of Academic Readiness standard setting technical report*. Retrieved March, 2014 from http://www.tea.state.tx.us/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=25769804117&libID=25769804117

Texas Education Agency (2013b). *State of Texas Assessments of Academic Readiness vertical scale technical report*. Retrieved March, 2014 from http://www.tea.state.tx.us/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=25769806053&libID=25769806056

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