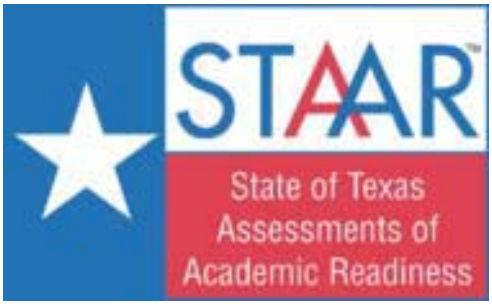


State of Texas Assessments of Academic Readiness (STAAR™) Assessments

Standard Setting Technical Report

March 15, 2013

Table of Contents

Chapter 1: The State of Texas Assessments of Academic Readiness (STAAR™)

This chapter provides an overview of the STAAR program and includes the following sections:

- Goals of the STAAR Program
- STAAR Curriculum Standards
- STAAR Performance Standards
- How STAAR Differs from TAKS

Goals of the STAAR Program

The 80th and 81st sessions of the Texas Legislature called for a new state assessment program to replace the Texas Assessment of Knowledge and Skills (TAKS), with the aim of continuing to use statewide student assessments to improve the state's education system. One of the state's goals in developing STAAR is that Texas will be among the top 10 states for graduating college-ready students by the 2019–2020 school year.

Toward this end, the Texas Education Agency (TEA), in collaboration with the Texas Higher Education Coordinating Board (THECB) and Texas educators, has developed STAAR to be a more rigorous assessment that provides the foundation for a new accountability system for Texas public education. STAAR is based on a new assessment model which includes the following.

- Performance expectations for STAAR were established so that graduating students are "postsecondary ready."
- The focus of student performance at high school shifted to 15 end-of-course (EOC) assessments. The 15 assessments, where appropriate, were linked to readiness for postsecondary endeavors, such as postsecondary education or career opportunities.
- The STAAR program was designed to be a comprehensive system, with curriculum and performance standards aligning with and linking back to elementary and middle school (grades 3–8) and projecting forward to postsecondary readiness.

The sections that follow provide a high-level description of how the curriculum and performance standards were determined for STAAR in order to meet the goals and requirements of the new assessment program.

STAAR Curriculum Standards

The curriculum assessed on STAAR is the state-mandated curriculum, the Texas Essential Knowledge and Skills (TEKS). These standards are designed to prepare students to succeed in postsecondary opportunities and to compete globally. However, consistent with a growing national consensus regarding the need to provide a more clearly articulated K–16 education program, STAAR focuses on fewer skills and addresses those skills in a deeper manner. By

focusing on the TEKS that are most critical to assess, STAAR measures the academic performance of students as they progress from elementary to middle to high school. While STAAR assessments at grades 3–8[1] address only those TEKS taught in the given subject and grade, the EOC assessments address only the TEKS for a given course, as opposed to the high school-level TAKS assessments, which addressed TEKS from multiple courses. Doing so strengthens the alignment between what is taught and what is tested for a given course of study.

Based on educator committee recommendations for each grade or course, TEA has identified a set of knowledge and skills from the TEKS that are eligible to be assessed. One subset of the TEKS, called readiness standards, is emphasized on the assessments. Other knowledge and skills are considered supporting standards and are assessed, although not emphasized.

Readiness standards have the following characteristics:

- They are essential for success in the current grade level or course.
- They are important for preparedness for the next grade level or course.
- They support postsecondary readiness.
- They necessitate in-depth instruction.
- They address broad and deep ideas.

Supporting standards have the following characteristics:

- Although introduced in the current grade or course, they may be emphasized in a subsequent grade or course.
- Although reinforced in the current grade or course, they may be emphasized in a previous grade or course.
- They play a role in preparing students for the next grade or course but not one that is central.
- They address more narrowly defined ideas.

Figure 1.1 shows the relative relationship between the readiness and supporting standards in the TEKS content standards and the readiness and supporting standards that are assessed each year. The STAAR assessment blueprints are designed so that a larger number of test items measure student expectations designated as readiness standards.

[1] Although the new science assessments for grades 5 and 8 continue to address TEKS from multiple grade levels, these tests will focus on the science TEKS for those respective grades. The science assessments at these two grades will emphasize the 5th- and 8th-grade curriculum standards that best prepare students for the next grade or course; in addition, these assessments will include curriculum standards from two lower grades (i.e., grades 3 and 4 or grades 6 and 7) that support students' success on future science assessments. In contrast, TAKS assessments uniformly addressed TEKS from multiple grade levels without any specific emphasis.

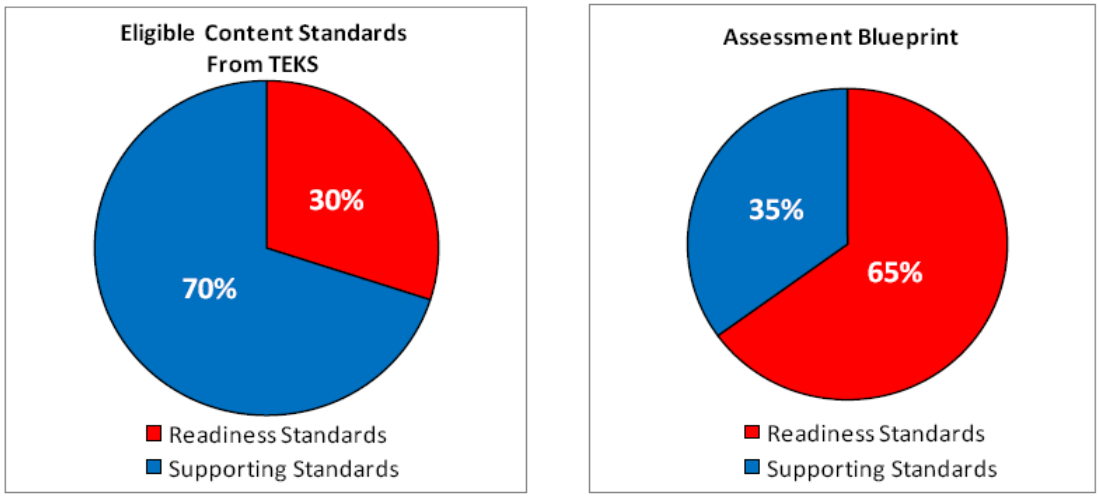


Figure 1.1: Readiness and Supporting Standards in the TEKS and Assessment Blueprint

TEA has also implemented a number of changes in the STAAR test design that serve to assess knowledge and skills in a deeper way.

- Tests contain a greater number of items that have a higher cognitive complexity level.
- Questions are developed to more closely match the cognitive complexity level evident in the TEKS.
- In reading, greater emphasis is given to critical analysis than to literal understanding.
- In writing, students are required to write two essays rather than one.
- In mathematics, science and social studies, process skills are assessed in context, not in isolation, which allows for a more integrated and authentic assessment of these content areas.
- In science and mathematics, the number of open-ended (griddable) questions has increased to allow students more opportunity to derive an answer independently.

STAAR Performance Standards

In addition to the new assessment design used for STAAR that focuses on fewer skills and that addresses those skills in a deeper manner, new performance standards had to be established for STAAR in order to satisfy legislative requirements for a new and more rigorous assessment system. The focus of this report is on the process used to establish the STAAR performance standards.

Performance Level Requirements

Federal statute requires any statewide assessment used for accountability (adequate yearly progress or AYP) purposes to include at least three achievement levels. In order to obtain at least three achievement levels, any STAAR assessment used for federal accountability needs to have at least two cut scores, or performance standards: one that distinguishes the “Level I” and “Level II” achievement levels and one that distinguishes the “Level II” and “Level III” achievement levels.

In addition, Texas Education Code (TEC) requires the establishment of specific performance standards on each STAAR assessment. For all STAAR assessments, there must be a cut score indicating satisfactory performance. For STAAR EOC assessments, a minimum score (within a reasonable range of the satisfactory cut score) must be established for use in determining whether a student's score on a particular EOC assessment may count toward his or her cumulative score in that content area. The cumulative score is used as part of a student's high school graduation requirements (see "Graduation Requirements" below). Also, performance standards indicating postsecondary or advanced-course readiness must be established for designated EOC assessments. Postsecondary-readiness standards are required for the STAAR Algebra II and English III assessments, while advanced-course readiness indicators are required for the Algebra I, English I, and English II assessments. Postsecondary-readiness standards may also be set at a later time for EOC assessments in science and social studies, depending on decisions based on the findings of the postsecondary-readiness feasibility study submitted to the legislature in December 2012. Details of the state legislative requirements can be found in Appendix 1.

STUDENT SUCCESS INITIATIVE

Enacted by the 76th Texas Legislature in 1999, the Student Success Initiative (SSI) grade advancement requirements apply to the STAAR reading and mathematics tests at grades 5 and 8. As specified by these requirements, a student may advance to the next grade level only by passing these tests or by the unanimous decision of his or her grade placement committee that the student is likely to perform at grade level after additional instruction. The goal of the SSI is to ensure that all students receive the instruction and support they need to be academically successful in reading and mathematics. Details of the state legislative requirements can be found in Appendix 1.

GRADUATION REQUIREMENTS

Texas Education Code specifies that, beginning with incoming grade 9 students in the 2011–2012 school year, testing requirements specific to the STAAR EOC assessments are used to determine eligibility for high school graduation. With STAAR, students may graduate through one of three programs: the Minimum High School Program (MHSP), the Recommended High School Program (RHSP), and the Distinguished Achievement Program (DAP). Table 1.1 outlines each high school program's course and assessment-related requirements, as required by the TEC.

Table 1.1: Graduation Requirements for High School Programs in Texas (starting in 2011–2012)

Program	Course Requirements* (pertaining to STAAR EOC)	Assessment-Related Requirements
Applies to all programs		o Student is required to achieve a cumulative score that is at least the product of the number of STAAR EOC assessments administered in a content area and the scale score that indicates Level II: Satisfactory Academic Performance. o Student must achieve a minimum score, which is set within a reasonable range of the satisfactory performance standard, in order for the score to count toward the student's cumulative score. o A student's cumulative score is determined using the student's highest score on each STAAR EOC assessment he or she is required to take for graduation purposes.
MHSP**	o Algebra I, geometry o Biology o English I, II, and III reading o English I, II, and III writing o U.S. history and either world geography or world history	o Student must be administered STAAR EOC assessments only for courses in which the student is enrolled and for which an EOC assessment is offered o Assessment scores only for courses specifically listed on the MHSP are required to count toward the cumulative score.
RHSP	o Algebra I, geometry, Algebra II o biology, chemistry, physics o English I, II, and III reading o English I, II, and III writing o world geography, world o history, and U.S. history	o Student must take all 15 STAAR EOC assessments. o In addition to the cumulative score requirements, a student must meet or exceed the Level II: Satisfactory Academic Performance standards for the STAAR English III reading, English III writing, and Algebra II assessments.
DAP	o Algebra I, geometry, Algebra II o biology, chemistry, physics o English I, II, and III reading o English I, II, and III writing o world geography, world o history, and U.S. history	o Student must take all 15 STAAR EOC assessments o In addition to the cumulative score requirements, a student must meet or exceed the Level III: Advanced Academic Performance standards that indicate postsecondary readiness for the STAAR English III reading, English III writing, and Algebra II assessments.

* These are the course requirements that pertain specifically to the STAAR EOC assessments. Students may be required to take additional courses under each graduation program.

** The specific curriculum and testing requirements for the minimum high school program in the mathematics, science, and social studies content areas may vary based on each student's course selection.

According to the requirements in Table 1.1, Texas high school students on the MHSP will need to take at least 11 EOC assessments, while students on the RHSP or DAP will need to take all 15 EOC assessments.

CUMULATIVE SCORE REQUIREMENT

Students receive test scores for each STAAR EOC assessment taken. A student's cumulative score is obtained by combining the individual test scores within each of the four foundation content areas (English reading/writing, mathematics, science, and social studies). For example, a student whose test scores in mathematics are 4200 for STAAR Algebra I, 3800 for STAAR geometry, and 4100 for STAAR Algebra II would have a cumulative score of 12100, the scores for all three mathematics assessments added together. For the score to count toward the student's cumulative score, he or she must achieve a minimum score, established at one conditional standard error of measurement (CSEM) below the satisfactory performance standard (See Chapter 8 for information about how the minimum score was set).

In order to graduate, students must reach or exceed their cumulative score target, which is based on the satisfactory performance standard for each content area. The specific cumulative score target for each student will vary depending on the student's graduation plan and when he or she started taking high school courses and the corresponding EOC assessments in Texas. As an illustration, consider the final scale score indicating satisfactory performance of 4000 for the STAAR EOC mathematics assessments. In this case, the cumulative score target for mathematics is 12000 for the RHSP and the DAP. If the hypothetical student in the previous example were on the RHSP or DAP, then he or she would have met the cumulative score target for the mathematics content area. If this student also meets his or her cumulative score target in each of the other foundation content areas (i.e., English reading/writing, science, and social studies), he or she would have satisfied the cumulative score requirements for high school graduation, as required by the TEC.

How STAAR Differs from TAKS

The STAAR assessment program differs from the current TAKS program in a number of significant ways.

- The STAAR assessment program has a stronger emphasis on academic rigor, both in terms of the number of tests that students need to take for graduation (11 to 15 in STAAR vs. four in TAKS) and the cognitive demands and level of skills needed to pass each assessment.
- The legislation requiring the new assessment program also focuses on the full spectrum of student performance. The goal is to make the STAAR program a comprehensive system, with curriculum and performance standards aligning and linking back to elementary and middle school (grades 3–8) and projecting forward to postsecondary readiness. Figure 1.2 provides a visual representation of this goal for the STAAR program.

- Empirical research studies are required to support the correlations or links between assessments in the same content area from elementary through high school. Such empirical linking studies are specifically required by the legislation for mathematics (i.e., grades 3–8 mathematics and Algebra I and II) and English (i.e., grades 3–8 reading and English I, II, and III reading).

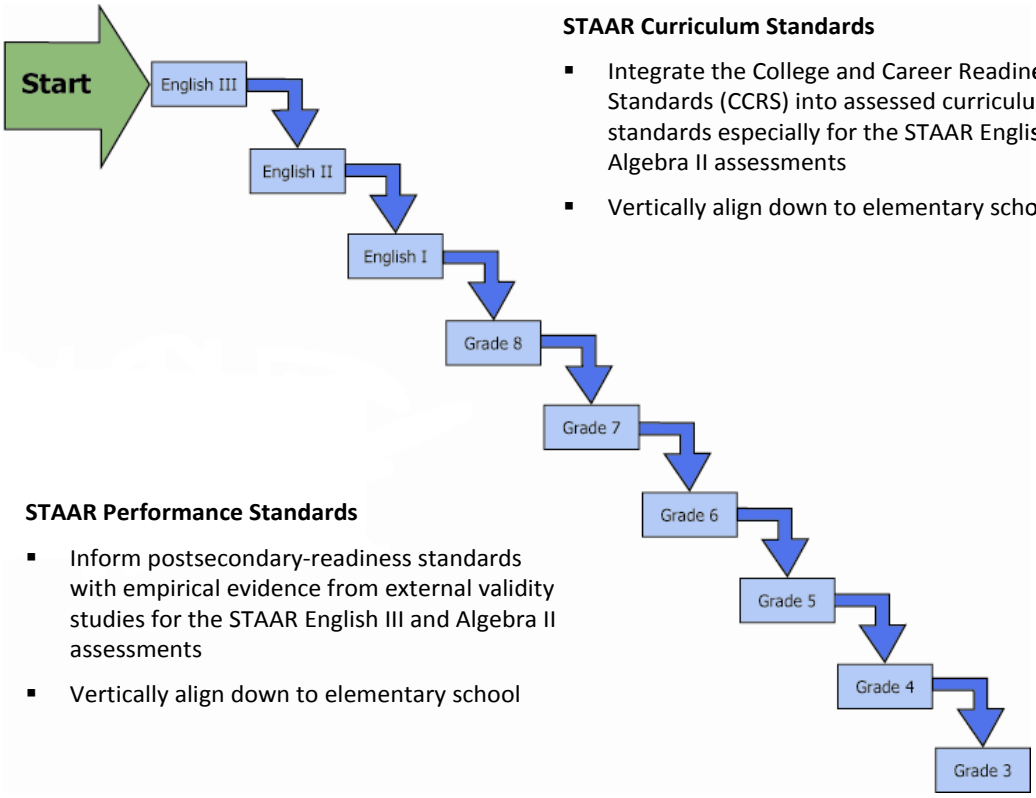


Figure 1.2: Vertical Alignment of Curriculum and Performance Standards for the STAAR Program

With these notable differences between the STAAR and TAKS programs, the process for setting performance standards was expanded beyond the process used for TAKS. Specific extensions to the standard-setting process are listed below.

- The STAAR standard-setting process took into account not only the assessed curriculum and content but also policy considerations and postsecondary readiness. Texas educators and content experts as well as policy experts and other stakeholders, such as those from the higher education and business communities, were part of the standard- setting process.
- Where practicable, scores on each STAAR assessment were empirically linked to scores on previous and successive assessments in the same content area. Satisfying a performance standard on one assessment helps establish how a student is expected to perform in a subsequent or advanced course and/or test in the content area and, in some cases, whether the student is secondary and postsecondary ready.

- Performance standards were required to have empirical evidence supporting that they mean what they are intended to mean. To meet this requirement, performance standards were externally validated by research studies that empirically correlate performance on the STAAR assessments with scores on other related measures or external assessments. (See Chapter 3 for information about the specific validity studies that were used to inform standard setting).

The next chapter provides more detail about the methodology and steps used to establish the STAAR performance standards

Chapter 2: Overview of the STAAR Standard-Setting Process

This chapter provides an overview of the STAAR standard-setting process and includes the following sections:

- Goals of Setting Performance Standards
- Evidence-Based Standard Setting
- The STAAR Standard-Setting Process

Goals of Setting Performance Standards

A critical aspect of any statewide testing program is the establishment of performance levels that provide a frame of reference for interpreting test scores. Once an assessment is administered, students, parents, educators, administrators, and policymakers want to know, in clear language, how students performed on that assessment. In general, by relating test performance directly to the student expectations expressed in the state curriculum in terms of what content and skills students are expected to demonstrate upon completion of each grade or course, performance standards describe the level of competence students are expected to exhibit.

Evidence-Based Standard Setting

As Texas implemented the STAAR program, which includes indicators of postsecondary readiness, TEA used a more evidence-based standard-setting approach (O'Malley, Keng, & Miles, 2012) than was used on TAKS. Standard setting for STAAR involved a process of combining considerations regarding policy, the TEKS content standards, educator knowledge about what students should know and be able to do, and information about how student performance on statewide assessments aligns with performance on other assessments. Standard-setting advisory panels composed of diverse groups of stakeholders considered the interaction of these elements for each STAAR assessment.

Figure 2.1 illustrates the critical elements of this evidence-based standard-setting approach used by Texas to establish the STAAR performance standards.

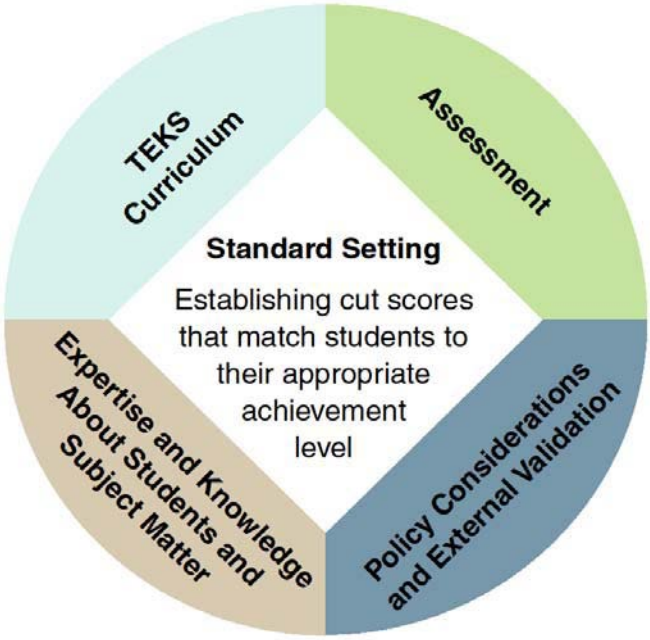


Figure 2.1: Critical Elements of the Evidence-Based Standard-Setting Approach

Each element of the evidence-based standard-setting approach as it relates to the STAAR assessments is described below.

- **TEKS Curriculum Standards**: The TEKS curriculum standards contain the content standards designed to prepare students to succeed in college and careers and to compete globally. They provide the underlying basis for several key components of the standard-setting process, including the performance labels, policy definitions, and specific performance level descriptors.
- **Assessment**: Each STAAR assessment has been developed to assess the knowledge and skills described in the TEKS curriculum standards. Each STAAR assessment is based on the student expectations and reporting categories specified in the STAAR assessed curriculum document and the STAAR test blueprint.
- **Policy Considerations and External Validation**: Research studies, which empirically correlate performance on the STAAR assessments with scores on other related measures or external assessments, were conducted and used to inform the standard- setting process. Stakeholders and experts with experience in educational policy and knowledge of the Texas assessment program considered the results of the research studies when making recommendations about reasonable ranges for setting performance standards.

- **Expertise and Knowledge about Students and Subject Matter**: Texas educators, including classroom teachers and curriculum specialists from elementary, secondary, and higher education, brought content knowledge and classroom experience to the standard-setting process. They played an integral role in developing the performance labels, policy definitions, and specific performance level descriptors and in recommending the performance standards.
- **Standard Setting**: Within the framework of evidence-based standard-setting, an established standard-setting method known as the bookmark method with external data (Ferrara, Lewis, Mercado, D'Brot, Barth, & Egan, 2011; Phillips, 2011) was used to recommend the cut scores, or performance standards.

The STAAR Standard-Setting Process

To fulfill legislative requirements, a nine-step process was followed in order to establish performance standards for STAAR assessments:

1. Conduct validity and linking studies
2. Develop performance labels and policy definitions
3. Develop grade/course specific performance level descriptors
4. Convene policy committee and 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

Tables 2.1 and 2.2 provide high-level descriptions and timelines for the steps in the STAAR EOC and 3–8 standard-setting process, respectively. Each step is described in detail in the remaining chapters of this report.

Table 2.1: Overview of the STAAR EOC Standard-Setting Process

Standard-Setting Step	Description	Timeline
1. Conduct validity and linking studies	External validity evidence was collected to inform standard setting and support interpretations of the performance standards. Scores on each assessment were linked to performance on other assessments in the same content area.	Studies started in spring 2009 and will continue throughout the program.
2. Develop performance labels and policy definitions	Committee convened jointly by TEA and THECB to recommend performance categories, performance category labels, and general policy definitions for each performance category.	September 2010
3. Develop grade/course specific performance level descriptors (PLDs)	Committees consisting primarily of educators developed performance level descriptors (PLDs) as an aligned system, describing a reasonable progression of skills within each content area (English, mathematics, science, and social studies).	November 2011
4. Convene policy committee	Committee considered policy implications of performance standards and empirical study results and made recommendations to identify reasonable ranges ("neighborhoods") for the cut scores.	February 1–2, 2012
5. Convene standard-setting committees	Committees consisting of K–12 educators and higher education faculty used the performance labels, policy definitions, PLDs, and neighborhoods set by the policy committee to recommend cut scores for each STAAR EOC assessment.	Mathematics and English: February 22–24, 2012 Science and Social Studies: February 29–March 2, 2012
6. Review performance standards for reasonableness	TEA and THECB reviewed the cut-score recommendations across content areas.	March 2012
7. Approve performance standards	The Commissioner of Education approved performance standards for satisfactory academic performance and advanced academic performance.*	April 2012
8. Implement performance standards	Performance standards were reported to students after the spring 2012 administration with phase-in standards applied.	May 2012
9. Review performance standards	Performance standards will be reviewed at least once every three years.	Fall 2014

* Minimum scores were also established empirically below the satisfactory and advanced academic performance standards and approved by the Commissioner of Education. Texas Success Initiative (TSI) exemption standards are under discussion by TEA and THECB.

Table 2.2: Overview of the STAAR 3–8 Standard-Setting Process

Standard-Setting Step	Description	Timeline
1. Conduct validity and linking studies	External validity evidence is collected to inform standard setting and support interpretations of the performance standards. Scores on each assessment are linked across grades to performance on other assessments in the same subject area.	Studies started in spring 2011 and will continue throughout the program.
2. Develop performance labels and policy definitions	Committee is convened jointly by the Texas Education Agency and the Texas Higher Education Coordinating Board to recommend performance categories, performance category labels, and general policy definitions for each performance category.	September 2010
3. Develop grade/subject specific performance level descriptors (PLDs)	Committees consisting primarily of educators develop performance level descriptors as an aligned system, describing a reasonable progression of skills within a subject area (reading, writing, mathematics, science, and social studies).	June 2012
4. Develop performance standard ranges	EOC performance standards and empirical study results are used to identify reasonable ranges (“neighborhoods”) for the cut scores for Levels II and III.	July 2012
5. Convene standard-setting committees	Committees consisting of K–12 educators use the performance labels, policy definitions, PLDs, and neighborhoods to recommend cut scores for each STAAR assessment.	October 2–12, 2012
6. Review performance standards for reasonableness	TEA reviews the cut-score recommendations across grades and subject areas.	October 2012
7. Approve performance standards	The Commissioner of Education approves performance standards.	December 2012
8. Implement performance standards	Performance standards are reported to students for the spring 2012 administration with phase-in standards applied.	January 2013
9. Review performance standards	Performance standards are reviewed at least once every three years.	Fall 2014

Chapter 3: Validity and Linking Studies

This chapter provides details about Step 1 of the nine-step STAAR standard-setting process, which focuses on conducting validity and linking studies. The sections in this chapter include

- Use of Empirical Evidence in Standard Setting
- Types of Empirical Studies
- Data Collection Design
- Analysis Methodologies
- STAAR EOC Empirical Studies
- STAAR 3–8 Empirical Studies
- Presenting Empirical Study Results
- Technical Issues and Caveats

Use of Empirical Evidence in Standard Setting

The STAAR assessment program is designed to be an aligned system of performance standards from grade 3 to high school. The STAAR performance standards are meant to provide indicators of the degree of preparedness for the next grade level, next course, or postsecondary readiness. Such standards relate 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, empirical evidence validates the use of those standards to describe academic content knowledge as well as the likelihood that students will meet future goals, such as success in the next grade level, next course, or postsecondary endeavors.

TEA in collaboration with THECB designed and implemented a systematic approach to incorporate empirical evidence into the STAAR standard-setting process. This approach was derived from an evidence-based standard-setting approach (Beimers, Way, McClarty, & Miles, 2012; O'Malley, Keng, & Miles, 2011). It blends components of several traditional standard-setting methods and was uniquely suited to fulfill the requirements of establishing performance standards for the STAAR assessments as required by state statute. The approach involved making use of the combined expertise of content specialists and measurement experts and included the following three steps:

1. **Determining the types of empirical studies to conduct**
   A framework was developed for determining empirical studies in order to gather a sufficient body of validity evidence.
2. **Developing data collection plans**
   Data were collected for the STAAR assessments and external tests to inform decision making and to meet timelines necessary to report student performance relative to the performance standards. Data were generally collected between 2009 and 2011 for STAAR EOC assessments and between 2011 and 2012 for STAAR 3–8 assessments.

3. **Selecting and executing appropriate analysis methods**
   Considerable planning and deliberation informed choices about statistical methodology. Each empirical study served a specific purpose during the standard-setting process, and each presented a unique set of requirements and considerations for quantitative analysis.

The next three sections of this chapter cover these steps in greater depth.

Types of Empirical Studies

The first step in incorporating empirical evidence into the STAAR standard-setting process focused on determining which studies to conduct. While some studies were specifically required based on legislation, others were discretionary. Additionally, it was important to balance having sufficient information to guide standard setting and having so much information that the data become difficult to interpret. If too many studies were presented, standard-setting panelists could be overwhelmed by the volume of empirical data. Thus, with the guidance of the Texas Technical Advisory Committee (TTAC) and input from the THECB, TEA systematically reviewed and selected empirical studies to collect an appropriate, but not overwhelming, body of validity evidence.

Master List of Potential Studies

To identify validity studies appropriate for standard-setting purposes, psychometric staff generated a master list of potential empirical studies. This list included studies that linked performance on a STAAR assessment with performance on other assessments within the Texas program (that is, internal studies). The list also included studies that linked performance on a STAAR assessment with performance on an external assessment or criterion (that is, external studies). Tables 3.1 and 3.2 provide examples of master lists specific to the STAAR English III and STAAR grade 8 reading assessments, respectively. Similar tables were generated for each STAAR assessment.

Table 3.1: Master List of Potential Empirical Studies for STAAR English III

Study Type	Studies empirically linking STAAR English III with…
Internal	• TAKS Grade 11 ELA • STAAR English I and STAAR English II • English III high school course grade
External	• College course grade • SAT • ACT • Advanced Placement (AP) • International Baccalaureate (IB) • SAT Subject Test • ACCUPLACER • Texas Higher Education Assessment (THEA) • COMPASS • ASSET • Program for International Student Assessment (PISA) • National Assessment of Educational Progress (NAEP) • WorkKeys • Armed Services Vocational Aptitude Battery (ASVAB)

Table 3.2: Master List of Potential Empirical Studies for STAAR Grade 8 Reading

Study Type	Studies empirically linking STAAR grade 8 reading with…
Internal	• TAKS grade 8 reading • STAAR English I reading • STAAR English I writing
External	• EXPLORE • ReadiStep • Program for International Student Assessment (PISA) • National Assessment of Educational Progress (NAEP) • Stanford Achievement Test – Tenth Edition (SAT-10) • Progress in International Reading Literacy Study (PIRLS) • Iowa Test of Basic Skills (ITBS)

STUDY SELECTION GUIDELINES

Next, potential empirical studies were described according to five key features: curricular relationships, legal requirements, data quality, types of performance standards, and visibility of the assessments. For each key feature, a set of selection guidelines was established. Through examination of each selection guideline, the value added by any given empirical study to the

standard-setting process was evaluated. Table 3.3 presents the guidelines specific to each of the five key features noted above.

Table 3.3: Guidelines for Selecting Empirical Studies

Key Feature	Selection Guidelines
Curricular Relationship	• There should be adequate access to the content of each assessment in the study to make the content comparisons. • There should be a reasonable amount of content overlap between the assessments in the study.
Legal Requirements	• A study specifically required by statute should be conducted. • Preference should be given to a study if it can help support the use of the STAAR EOC program as part of the graduation testing requirement.
Data Quality	• Studies with the following data characteristics are preferred: o Student-level data o Operational test data o Motivated data (that is, derived from high-stakes tests) o No additional data collection needed o Minimal time lapse between when the STAAR and external assessments are taken • Data from the assessments should be available in time to conduct the study for standard setting.
Type of Performance Standards	• It would be preferable to have at least one study that informs each cut score. • It would be preferable to have consistency of studies available across content areas (for example, mathematics and English). • Preference should be given to studies that can serve multiple purposes to avoid redundancy in analyses.
Visibility of Assessments	• The external assessment should be taken by students in Texas. • The external assessment should have national or international prominence. • The study should provide evidence about the rigor of STAAR. • The external assessments should be used in Texas to determine college readiness and/or placement. • Preference should be given to studies incorporating tests taken by special populations (for example, special education, English language learners, etc.).

SELECTION OF STUDIES

A subset of studies that would be most useful and informative for the STAAR standard settings was chosen based on the guidelines in Table 3.3. The list of potential studies and the process for selecting the studies were reviewed by the TTAC in June 2010 for STAAR EOC and STAAR 3–8. The TTAC agreed with the approach and provided suggestions to TEA that helped refine the selection process. In spring 2012, additional external studies were considered for STAAR 3–8

based on results from the STAAR EOC performance standards and availability of external data. The final set of empirical studies selected and conducted to support standard setting may be grouped into seven categories:

1. *Linking studies*, which link performance across assessments within content areas in the STAAR program (for example, Algebra I and Algebra II)
2. *STAAR-to-TAKS comparison studies*, which link performance on STAAR assessments to performance on TAKS assessments
3. *Grade correlation studies*, which link performance on STAAR EOC assessments to high school course grades
4. *External validity studies*, which link performance on STAAR assessments to external measures (specifically, SAT, ReadiStep, ACT, EXPLORE, THEA, and ACCUPLACER)
5. *NAEP and PISA comparisons*, which compare national and international assessment data to STAAR performance
6. *College students taking STAAR*, which link performance on STAAR EOC assessments to college course grades
7. *Vertical scale studies,* which allow the comparison of student performance across grades within a content area for grades 3–8 reading and mathematics

Appendix 2 provides a more detailed description of each of these studies relative to their use in the STAAR standard-setting process. It is important to note that the studies chosen were identified and conducted to support the initial STAAR standard-setting process. Texas legislation requires the review of performance standards at least once every three years (Step 9 in the STAAR standard-setting process). The framework for selecting empirical validity studies that informed the initial standard setting has also been used for identifying potential studies for standards review. Refer to Chapter 11 for additional details about the plans for standards review in the STAAR program.

Data Collection Design

Three data collection designs were implemented in order to conduct the empirical studies for the STAAR assessments: single-group design, coarsened exact matching, and common-item non-equivalent groups design. When establishing links between two tests, it is preferable to obtain scores on both tests from a common sample of students. This is known as a single-group design because all data related to a pair of linked tests are collected from one group of students who took both assessments. When a single-group design is not possible, a matching methodology known as coarsened exact matching (CEM, Iacus, King, & Porro, 2011) can be used to create a set of matched students. This matched sample is meant to imitate a single group. The CEM procedure matches the two student groups based on characteristics statistically associated with both tests. The characteristics may include gender, ethnicity, socioeconomic status, and an academic achievement composite based on assessments that both groups of students have taken. A third data collection design, common-item non-equivalent groups design, may be appropriate when a subset of items from one test is included in the administration of the other test to be linked (Kolen & Brennan, 2004). This design allows

items on both tests to be placed on the same scale. The following two sections discuss the data collection designs for the STAAR EOC studies and the STAAR 3–8 studies.

Analysis Methodologies

Several linking methodologies were used to analyze the collected data. The methods can be classified into three categories: equipercentile linking, regression-based methods, and item response theory methods. The linking methods are briefly introduced in this section. Detailed analytic steps and statistical models specific to each method are provided in Appendix 2. Applications of the various analysis methods for each study were reviewed with the TTAC during the October 2009, February 2010, June 2010, August 2011, March 2012, and September 2012 TTAC meetings. Table 3.4 lists the analysis methods applied for each of the STAAR empirical studies.

Table 3.4: Analysis Methods for STAAR Empirical Studies[2]

Empirical Study	STAAR EOC Methods Applied	STAAR 3–8 Methods Applied
Linking studies	Regression-Based Linking	Regression-Based Linking
STAAR-to-TAKS comparison studies	Equipercentile Linking	Equipercentile Linking Item Response Theory
Grade-correlation studies	Regression-Based Linking	N/A
External validity studies	Regression-Based Linking	Regression-Based Linking
College students taking STAAR	Regression-Based Linking	N/A
Vertical scale studies	N/A	Item Response Theory

EQUIPERCENTILE LINKING

The equipercentile linking method, which was developed to link SAT scores to ACT scores (Pommerich, Hanson, Harris & Sconing, 2004; Dorans, Lyu, Pommerich, & Houston, 1997), was used to conduct the STAAR-to-TAKS comparison studies. This linking method is appropriate when looking for scores on one assessment that are *equivalent* to scores on the linked assessment. The equipercentile method produces concordance tables through which equivalent TAKS scores may be identified on the STAAR scales. In the case of the STAAR-to-TAKS comparison studies, a concordance table that related scores on STAAR to those on TAKS was necessary to evaluate claims about the rigor of the STAAR performance standards relative to the rigor of the TAKS performance standards.

[2] For comparisons with NAEP and PISA, no empirical linking studies were conducted because no student-level data were available for these assessments. For NAEP, state- and national-level impact data were obtained directly from the most recent (2002, 2007, 2009, and 2011) administrations in each content area. For PISA, results based on established comparisons between the PISA scale and the ACT scales were considered.

REGRESSION-BASED LINKING

In cases where an empirical link between two assessments was needed but no assumptions about score equivalency were made, regression-based approaches could be applied. Empirical correlations were calculated for each empirical study to gauge the appropriateness of regression-based linking. In each case, the linear relationship between the two linked tests was sufficiently strong that a score on a STAAR assessment could be used to predict a score on an external test. Two types of regression-based approaches were used: logistic regression and ordinary least square (OLS) regression. Logistic regression provided the estimated probability that a test taker would achieve a certain level of performance on an external assessment, conditional on STAAR performance. OLS regression provided the estimated mean score on an external assessment, conditional on STAAR performance.

ITEM RESPONSE THEORY

When the data-collection design is based on common-item non-equivalent groups and items from one test are embedded in another test, item response theory places test items and measures of student proficiency on the same scale. The relationship between the two tests is determined based on the underlying item response theory scale. For the STAAR-to-TAKS comparison studies in grades 3–8, the TAKS Met Standard performance standards were identified on the STAAR assessments. In addition, the linking study based on the vertical scale analyses used item response theory to determine the relationship between tests in adjacent grades for reading and mathematics, which helped to align performance standards across grades for STAAR 3–8.

STAAR EOC Empirical Studies

For most of the empirical studies that informed the initial STAAR EOC standard-setting process, single-group designs were available. Data were gathered beginning in 2009 for the EOC assessments and from the 2010 and 2011 administrations of external tests (SAT, ACT, THEA, and ACCUPLACER). Data informing the STAAR-to-TAKS comparison studies, grade correlation studies, and college students taking STAAR studies were collected in 2011. The NAEP study relied on impact data comparisons rather than empirical links; results from both the 2009 and 2011 NAEP administrations were incorporated. Finally, PISA links relied on established comparisons between that assessment's scale and ACT scales, so no additional data collection was required.

In a few cases, single-group data collection designs were not feasible. The STAAR EOC linking studies compared STAAR EOC scores from consecutive courses within the same academic content area. These studies relied on tests administered sequentially (for example, English I reading to English II reading to English III reading). Table 3.5 shows the data collection schedule for all STAAR EOC linking studies used in the initial standard-setting process before the spring 2012 administration.

Table 3.5: Data Collection Schedule for STAAR EOC Linking Studies

Content Area	Spring/Fall 2009	Spring 2010	Spring 2011
Mathematics	Algebra I	Geometry	Algebra II
English (Reading and Writing)		English I	English II English III

As shown in Table 3.5, it was possible to implement a single-group design for the STAAR EOC mathematics assessments and the STAAR English I and II assessments by collecting test scores from a cohort of students longitudinally, beginning with spring/fall 2009 for mathematics and spring 2010 for English I and II. However, STAAR English II and English III were both administered (as field tests) for the first time in spring 2011. Because Texas students generally do not take these two courses in the same school year, it was not possible to collect pairs of scores from a single cohort of students who took both English II and English III before the initial standard setting. CEM was used to create a set of matched students from the spring 2011 English II and English III testers. The CEM procedure matched the two student groups based on characteristics statistically associated with scores on both English II and English III. The characteristics used to match the two groups included gender, ethnicity, socioeconomic status, and an academic achievement composite based on assessments that both groups of student were required to take.

STAAR 3–8 Empirical Studies

Several data-collection designs were used for the empirical studies for STAAR 3–8. For some studies, more than one data-collection design was available, which required evaluating the quality of the data and the strength of the relationship between the linked assessments. Appendix 4 provides more detail regarding the data-selection decisions, the quality of the linking studies, and the analysis method for each empirical study.

STAAR 3–8 Empirical Links with EOC

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. The data-collection design consisted of both single-group design and coarsened exact matching. Logistic regression analyses provided the probability of attaining a particular score on the STAAR EOC assessments given a student's performance on the STAAR grade 8 assessments and the grade 7 writing assessment.

STAAR 3–8 Empirical Links Across Grades

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. The data-collection design consisted of both single-group design and coarsened exact matching. Logistic regression analyses provided the probability of attaining a particular

score on a subsequent test within a content area given a student's performance on a STAAR assessment.

STAAR 3–8 External Validity Studies

The STAAR grade 8 assessments and the grade 7 STAAR writing assessment were linked to external measures—EXPLORE and ReadiStep—which are linked to ACT and SAT, respectively. Data collection was based on single-group design. Logistic regression analyses provided the probability of attaining a particular score on the external measures given a student's performance on the STAAR grade 8 assessments and the grade 7 writing assessment.

STAAR 3–8 Vertical Scale Studies

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. Data collection design followed a common-item non-equivalent groups design in which students took on-grade-level items and off-grade-level items from adjacent grade levels for reading and mathematics. Item response theory was used to estimate the vertical scales. 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.

STAAR-to-TAKS Comparisons

Studies compared performance on STAAR to performance on TAKS in order to ensure that the performance standards for STAAR are more rigorous than TAKS performance standards. Data collection included single-group design and common-item non-equivalent groups design. Equipercentile equating and item response theory were used to attain the TAKS Met Standard performance level on the STAAR assessments. The empirical result was evaluated with respect to trends in TAKS impact data and the impact data for the STAAR 2012 assessments.

Using the analysis methods listed in Table 3.4, the STAAR empirical studies were conducted over the summer and fall of 2011 (for STAAR EOC) and summer 2012 (for STAAR 3–8) as data were collected and made available. Study results were summarized, reviewed, and presented to a variety of audiences. The next section describes how the empirical study results were presented to various stakeholders and committees involved in the STAAR standard-setting process.

Presenting Empirical Study Results

Results from empirical studies provide value to the STAAR standard-setting process only if they can be communicated clearly and accurately to the intended users of the study results. Therefore, TEA and THECB carefully considered how to present the validity and linking study results to various audiences, particularly those who are non-technical. Several approaches to consolidating study results for the purposes of sharing with subsequent committees were considered. These approaches were presented both to the TTAC on several occasions (October 2009, March 2011, and August 2011) and to individual TTAC members during the months leading up to the policy and standard-setting committees. The TTAC provided valuable feedback on the approaches that were incorporated into the final presentation of the study results. In

general, four main approaches were used to communicate the empirical study results during the STAAR standard setting process:

1. Empirical number lines
2. Quality summary and study profiles
3. Likelihood tables
4. Vertical scale graphs

Each of these approaches is described in the following subsections.

Presenting STAAR EOC Results

EMPIRICAL NUMBER LINES

A horizontal number line was used to show how various empirical study results fall relative to one another for a particular STAAR EOC assessment. The values displayed on the number line were the percentage of students (based on performance on the spring 2011 administration) who scored at or above this point on the STAAR EOC assessment of interest. This scale metric was chosen so that users could easily see the percentage of students that would meet or exceed a cut score if it were strictly aligned with the result of a particular study. One of the main impetuses for this approach was to reduce the amount of numbers displayed so that users could focus on the relative positions and implications of the various study results. This approach, therefore, was instrumental in the communication of the reasonable ranges, or neighborhoods, for each performance standard considered by the policy committee (see Chapter 6).

Figure 3.1 shows the empirical number line for the STAAR Algebra II assessment. As a specific example, consider the highlighted call-out box labeled “ACT math CR (college readiness) benchmark” on the number line. The ACT mathematics test was empirically linked to STAAR Algebra II via a single-group design. Logistic regression analysis indicated that students with a Rasch-based Algebra II ability estimate (θ) of 0.1 would have an approximately 50% chance of meeting the college-readiness benchmark on the ACT mathematics assessment (see “A Note about External Benchmarks” below). Because the Rasch scale is unfamiliar to most educators and practitioners and scale scores could not be established until the conclusion of the standard-setting process, this value was converted to impact data: 27% of Algebra II testers have a θ estimate of at least 0.1. In effect, that percentage — 27% — served as the result of the Algebra II–ACT mathematics external validity study. A similar procedure was used to derive call-out boxes for all the empirical studies involving the STAAR Algebra II assessment.

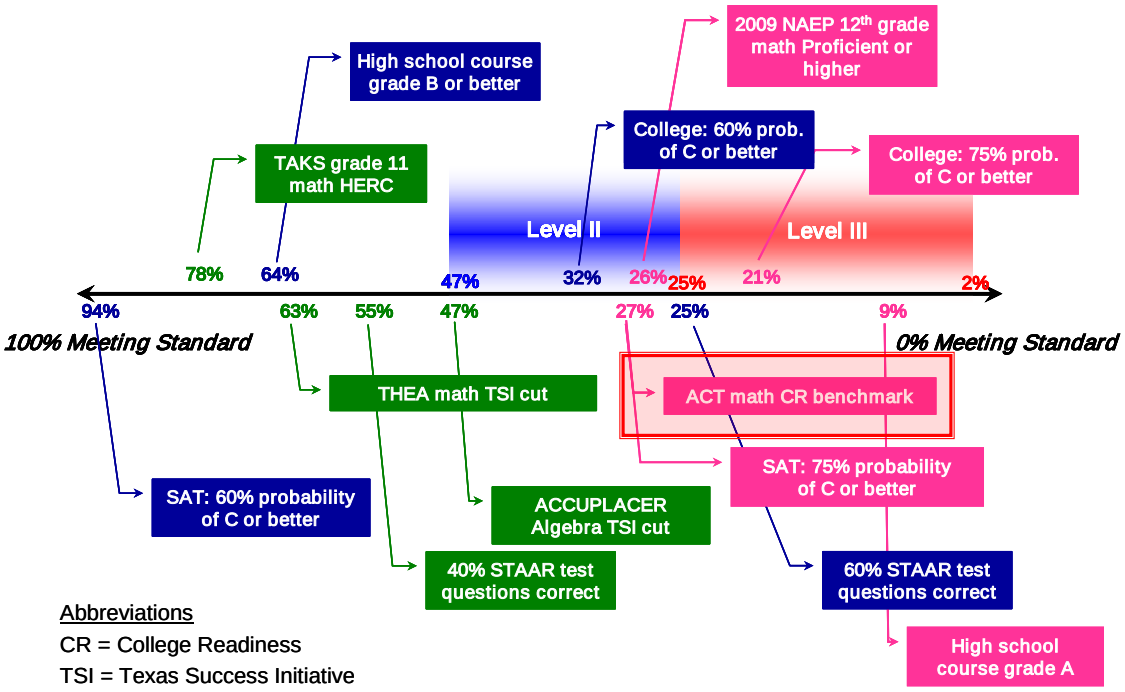


Figure 3.1: Empirical Number Line for STAAR Algebra II

In presenting the study results, the empirical number line was a dynamic display, in which each call-out box appeared one at a time. Presenting them in this progressive manner allowed each study to receive full attention from the audience without distracting or overwhelming panelists with the other study results. The color scheme and the order of presentation of the studies were also carefully chosen. Studies meant to be well below the satisfactory (Level II) standard were colored green and shown first. Studies shaded blue were next and provided information about where it would be reasonable to set the Level II cut score. Finally, studies that informed where it would be reasonable to set the Level III cut score were displayed in pink and presented last. Chapter 6 provides additional information about the development guidelines of the reasonable ranges (or neighborhoods) for each of the STAAR EOC cut scores.

A Note about External Benchmarks. Many of the studies shown on the empirical number line were presented with reference to established benchmarks on the external tests. This was illustrated above with the college-readiness benchmark on the ACT mathematics test. ACT, SAT, THEA, and ACCUPLACER assessments each have established cuts scores indicating that students either are likely to succeed in college (ACT and SAT) or do not require remediation before beginning postsecondary coursework (THEA and ACCUPLACER). As part of the analysis, points along the STAAR scales indicating that students would be at least 50% likely to meet or exceed those external benchmarks were estimated. Those reference points were used in locating the call-out boxes for each study on the number line. To clarify the benchmarks examined, Table 3.6 provides the external tests to which each STAAR EOC assessment was linked, along with the cut scores examined for those linked tests.

Table 3.6: Measures and Benchmarks Linked to STAAR EOC Assessments

STAAR Assessment	Linked Test	Reference Point
Algebra II	ACT Mathematics	22
Algebra II	SAT Mathematics	390, 510[1]
Algebra II	THEA Mathematics	230[2]
Algebra II	ACCUPLACER Algebra	63[2]
English III Reading	ACT Reading	21
English III Reading	SAT Critical Reading	340, 390[1]
English III Reading	THEA Reading	230[2]
English III Reading	ACCUPLACER Reading	78[2]
English III Writing	ACT English	18
English III Writing	SAT Writing	280, 310[1]
English III Writing	THEA Writing	220[2]
English III Writing	ACCUPLACER Sentence Skills	80[2]
English III Writing	ACCUPLACER Written Essay	6[2]
Biology	ACT Science	24
Biology	SAT Mathematics	410, 470[1]
Chemistry	ACT Science	24
Chemistry	SAT Mathematics	420, 470[1]
Physics	ACT Science	24
Physics	SAT Mathematics	410, 480[1]
World Geography	ACT Reading	21
World Geography	SAT Critical Reading	320, 390[1]
U.S. History	ACT Reading	21
U.S. History	SAT Critical Reading	320, 390[1]

[1] Two links were provided to SAT scales; the lower score represents a 60% chance of earning a C or better in a corresponding college course, while the higher score represents a 75% chance of the same outcome. Refer to Chapter 6 for the rationale for these probability values.

[2] THEA and ACCUPLACER benchmarks match cut scores established for the Texas Success Initiative (TSI). Refer to Chapter 6 for more information about TSI.

QUALITY SUMMARY AND STUDY PROFILES

Given the number of separate studies conducted, it was important that users of the empirical study results be able to evaluate the quality of the data underlying the estimates they interpreted. The source data varied according to five identified dimensions:

1. **Motivation** of students taking each assessment
2. **Representativeness** of the students in the study sample
3. **Sample size**
4. **Correlation** between scores on linked assessments
5. Degree of **content overlap** between those assessments

TEA rated each empirical study according to these five dimensions. In addition, an overall rating was produced. The overall rating was calculated by taking a weighted average of the dimension ratings, where statistical correlation was double-weighted to adequately emphasize the effect of prediction error in regression-based methods. The ratings for the STAAR EOC external validity studies were summarized in a quality summary table to enable the policy committee to evaluate the quality of the studies and enable easy comparisons across the studies. Appendix 3 provides the STAAR EOC quality summary table.

Furthermore, STAAR EOC study profiles were produced to provide more detailed information about the purpose and characteristics of each empirical study. The study profiles were constructed following the same framework as the quality summary table, but in the study profiles the rationale underpinning each dimension's rating was articulated in greater detail. Additionally, each profile included information about the assessments used to construct links, such as test length, item formats, time limits, frequency of administration, and the performance standards established for the tests. Individual study profiles covering each external validity study and each *college students take STAAR* study can be found on the TEA website at http://www.tea.state.tx.us/staar/vldstd.aspx.

Although study profiles were crafted to display comparisons between STAAR EOC and external measures (such as SAT and ACT), additional documentation was provided covering STAAR-to-TAKS comparison studies, grade correlation studies, and linking studies. In these documents, sample sizes, correlations, and results along with general descriptions of each study's purpose are provided. The documentation for STAAR EOC studies is available at the above link on the TEA website. Additionally, the NAEP impact data that were presented to both policy committee members and standard-setting committee panelists are also available at the above link on the TEA website.

LIKELIHOOD TABLES

Likelihood tables were used during the standard-setting committee meetings (see Chapter 7) to provide panelists feedback on the implications of their recommended cut scores relative to various benchmarks. Table 3.7 gives an example of a likelihood table.

Table 3.7: Example Likelihood Table

Performance Standard	Level II	Level III
Probability of a C or higher in an entry-level college course	67%	91%
Projected SAT score	472	609
Projected ACT score	21	28

The likelihood information shown in Table 3.7 is based on the committee's recommended cut scores after a particular round of judgment. The table includes mean SAT scores, mean ACT

scores, and the average likelihood that students in each performance level would succeed in the next course in high school or college. "Success" in this context was defined as either achieving the equivalent performance level in a subsequent high school course or passing an entry-level college course in the same content area. A low projected likelihood of success (e.g., less than 30%) would suggest that the recommended cut score was low relative to the passing standard at the subsequent level. To contextualize projected SAT and ACT scores, facilitators provided panelists with reference points that included average SAT and ACT scores of enrolled college students nationally and in Texas. Table 3.8 provides an example of the reference points given to panelists. Refer to Chapter 7 for more information about the standard-setting committees and the meeting proceedings.

Table 3.8: Example Reference Points

Reference Point	SAT	ACT
National Average	497	21.3
Texas State Average	479	20.7

Presenting STAAR 3–8 Results

Vertical Scale Graphs

Similar data presentations were developed for STAAR 3–8 as were used for STAAR EOC. Table 3.9 gives an example of a likelihood table for the STAAR grade 8 mathematics standard-setting committee meeting, which provided panelists feedback on the implications of their previous round's recommended cut scores (median page number) for Level II and Level III relative to ReadiStep and EXPLORE. In this example, a typical student in the Level III performance category has a 76% probability of reaching the EXPLORE benchmark based on the previous round's recommendation for Level III.

Table 3.9: Example of Likelihood Table for Linking STAAR Grade 8 Mathematics to ReadiStep and EXPLORE

Performance Standard	Level II	Level III
Borderline Student		
Probability of reaching the EXPLORE benchmark	26	71
Probability of reaching the READISTEP benchmark	47	90
Typical Student		
Probability of reaching the EXPLORE benchmark	43	76
Probability of reaching the READISTEP benchmark	69	92

In addition, STAAR 3–8 vertical scale results were presented. The vertical scales for reading and mathematics empirically link student performance on STAAR 3–8 assessments within the same

subject area. Because student performance on a vertical scale can be compared from grade to grade in order to gauge academic progress in mathematics or reading across time, the vertical scale was used to evaluate the alignment of performance standards across assessments. The reasonable ranges for performance standards were prepared using the alignment of the vertical scale across grades. In addition, the vertical scale allowed standard-setting panelists to consider the progression of performance standards across grades for their specific grade in relation to previously recommended performance standards for higher grades. For example, the grade 5 mathematics committee considered the recommended performance standards for grades 6, 7, and 8 mathematics as one piece of information in recommending the grade 5 performance standards. Figure 3.2 presents an example of a vertical scale graph provided to members of the standard-setting committee as feedback data so that they could evaluate their judgments relative to where prior committees had recommended performance standards on the upper-grade-level assessments.

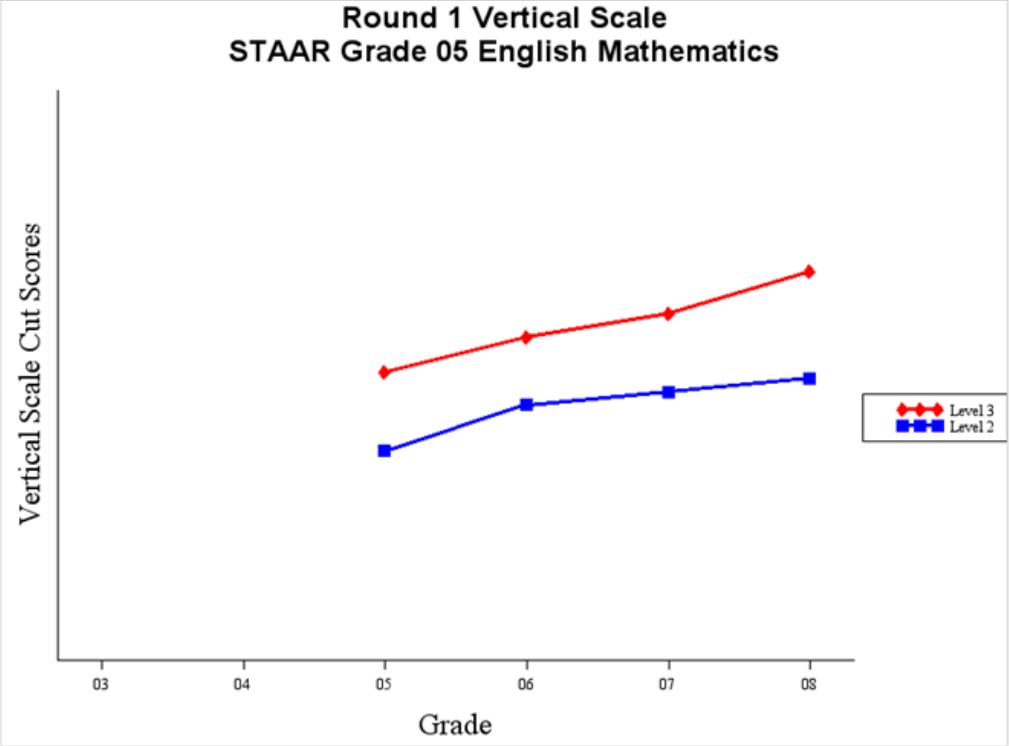


Figure 3.2: Example of Vertical Scale Feedback Data for STAAR Grade 5 Mathematics

Technical Issues and Caveats

The most important caveat related to empirical validity studies is the impact of student motivation on assessment performance. All STAAR EOC data used for the validity studies were collected before or during 2011. This data collection schedule was necessary if evidence-based standards were to be established before the first high-stakes administrations of STAAR in spring 2012. However, test administrations before or during 2011 did not carry state-imposed consequences for students who participated. Low-stakes testing scenarios may reduce

students' motivation to participate in assessments with the same level of effort they would under high-stakes conditions. As a result, with unmotivated data, estimated pass rates at any given score point will likely be artificially low. It is also reasonable to suspect that statistical correlations between unmotivated tests (such as STAAR) and motivated external measures (such as SAT) will be lower than correlations between pairs of motivated scores.

A close examination of the STAAR data combined with the use of common data visualization techniques such as scatter plots, stem-and-leaf plots, box plots, and identification of outlying data points were helpful in identifying unmotivated scores. For example, by examining response patterns on written composition sections of English assessments, it was possible to identify students submitting unmotivated responses and evaluate the impact of their scores on aggregate study results. The potential effect of unmotivated responses was emphasized when presenting impact data (i.e., projected pass rates) to the policy committee and standard-setting committees. The policy committee was also provided with content-area-specific estimates of how pass rates could change under more motivated conditions.

A caveat for the STAAR 3–8 assessments is the lack of cohort data in conducting the empirical studies in spring 2012. Only one administration was completed when the analyses were required. As students progress through the STAAR system, additional data from cohorts will be available, thereby allowing additional single-group designs.

The sections in this chapter have detailed the data-collection designs and statistical methodologies that underpin the validity studies. The results of the empirical validity and linking studies were presented to the committees of educators and policymakers responsible for establishing STAAR performance standards. Refer to Chapters 6 and 7 of this report for more information about the results of these specific empirical studies.

Chapter 4: Performance Labels and Policy Definitions

This chapter provides details about Step 2 of the nine-step STAAR standard-setting process, which focused on developing performance labels and policy definitions. The sections in this chapter include the following:

- Performance Descriptor Advisory Committee Meeting Purpose
- PDAC Committee Composition
- PDAC Meeting Proceedings
- Outcome of the PDAC

Performance Descriptor Advisory Committee Meeting Purpose

TEA, in cooperation with THECB, convened a Performance Descriptor Advisory Committee (PDAC) on September 30–October 1, 2010, to recommend performance labels and policy definitions for the performance standards of the STAAR program. The purpose of the performance labels and policy definitions is to describe the general level of knowledge and skills evident at each performance level across all content areas and grades/courses. During the standard-setting committee meetings, these labels and definitions provided the panelists with a consistent baseline as they developed recommendations for the cut scores associated with each performance standard.

The Commissioner of Education charged the panelists to:

1. assume that the state assessment system will be implemented under current federal and state statute, both of which require a minimum of three performance levels;
2. reach consensus on recommendations for the names of the performance labels (categories of performance) for student achievement on the assessments (general, modified, and alternate); and
3. make recommendations for key words and phrases to be used in drafting the policy definitions that will define student performance within each category.

In addition, to help them make recommendations, panelists were given the following preliminary guidelines describing what effective policy definitions are.

- Policy definitions should communicate the degree to which students demonstrate knowledge and skills but should be generalizable across content areas and grades/courses.
- Policy definitions should be succinct and clear to the intended audience: schools, parents, students, and the general public.
- Policy definitions should be accurate descriptions focused on students who perform in the middle of the category and take into account the range of student performance within each category.

- Policy definitions should focus on student performance demonstrated on the assessments, not on other student characteristics.
- Policy definitions should include information about a student's readiness for the next grade/course.
- Policy definitions for the modified and alternate assessments will be different from definitions for the general assessments in that they take into account the unique needs of students with disabilities who take the modified or alternate assessments.

PDAC Committee Composition

Dr. Gregory Cizek, a professor in Educational Measurement and Evaluation at the University of North Carolina facilitated the meeting. The committee consisted of 26 panelists who were selected to represent the diversity of stakeholders in public education and higher education in Texas. The PDAC panelists' names, positions, and affiliations are listed in Table 4.1.

Table 4.1: Performance Descriptor Advisory Committee Panelists

Panelist Name	Panelist Position and Affiliation
Dana Bedden	Superintendent, Irving ISD
Reece Blincoe	Superintendent, Brownwood ISD
Bobby Blount	Director, Vice-Chair of Bylaws, Texas Association of School Boards
Von Byer	Committee Director, Senate Education Committee
Jesus Chavez	Superintendent, Round Rock ISD
Patti Clapp	Executive Director, Greater Dallas Chamber of Commerce
David Dunn	Executive Director, Texas Charter Schools Association
Andrew Erben	President, Texas Institute for Education Reform
Dora Garcia	Teacher, Los Fresnos CISD
Julie Harker	Public Education Advisor, Office of the Governor
Troy Johnson	Associate Vice President, University of North Texas
Sandy Kress	Partner, Akin, Grump, Strauss, Hauer, and Feld
Russell Lowery-Hart	Vice President of Academic Affairs, Amarillo College
Donna Newman	Executive Director of Middle School Performance, Hays CISD
Esmeralda Perez-Gonzalez	Teacher, Hays CISD
Anne Poplin	Director, ESC, Region IX
Richard Rhodes	President, El Paso Community College
Todd Rogers	Principal, Northwest ISD
Rod Schroder	President, Texas School Alliance
Jeri Stone	Executive Director, Texas Classroom Teachers Association
Tom Torkelson	Chief Executive Officer, IDEA Public Schools
Rod Townsend	President, Texas Association of School Administrators
Maria Trejo	Director of Curriculum and Instruction, Cypress-Fairbanks ISD
Gabriel Trujillo	Principal, Duncanville ISD
Lori Vetters	Chairperson, Pre-K Committee, Greater House Partnership
Jenna Watts	Policy Director, House Public Education Committee

To facilitate discussion during the meeting, committee members were divided into four groups, each of which represented a cross-section of panelists from K–12 education, higher education, special populations, business, and Texas state government. A selection of committee members from the PDAC were invited to attend subsequent meetings throughout the standard -setting process.

PDAC Meeting Proceedings

During the two-day meeting, the committee was led through a six-step process to develop recommendations for the performance labels and policy definitions:

1. Brainstorm key words and phrases to be used in developing the policy definitions.
2. Share recommendations for key words and phrases.
3. Reach consensus on recommendations for key words and phrases to be used in developing the policy definitions.
4. Brainstorm performance labels for each of the performance categories.
5. Share recommendations for performance labels.
6. Reach consensus on recommendations for performance labels.

Table 4.2 shows the agenda for the PDAC meeting.

Table 4.2: PDAC Meeting Agenda

Day 1	• Welcome and Introductions • Purpose, Goals, and Overview of Agenda • Overview of the STAAR Program, Comparison of TAKS and STAAR, Legislative Requirements, College Readiness • Graduation Plans and Performance on Assessments • Content Overview, Increased Rigor of Assessments, Alignment of Content Standards • Overview of the Standard-Setting Process, Standard-Setting Timeline, Alignment of Performance Standards • Performance Category Labels and Policy Definitions Overview • Small-Group Discussion of Key Words and Phrases for Policy Definitions (Step 1) • Recommendations from Small-Group Discussions (Step 2)
Day 2	• Review Recommendations from Small-Group Discussions on Day 1 • Reach Consensus on Recommendations for Key Words/Phrases for Policy Definitions (Step 3) • Performance Category Labels Overview • Small-Group Discussion of Performance Category Labels (Step 4) • Recommendations from Small-Group Discussions (Step 5) • Panel Reaches Consensus on Recommendations from Performance Category Labels (Step 6) • Review Panel Recommendations for Performance Category Labels and Key Words/Phrases • Additional Feedback/Recommendations from Committee • Concluding Remarks

A description of each of the six steps in the process is provided below.

STEP 1: BRAINSTORM KEY WORDS AND PHRASES TO BE USED IN DEVELOPING THE POLICY DEFINITIONS

Preliminary guidelines for developing the policy definitions were shared, the charges to the PDAC from the Commissioner of Education were read, and the following guiding principles were reviewed.

1. There will be three performance levels, each with a different label.
2. The labels should be different from the current TAKS labels.
3. The standards at each performance level will be linked from grade to grade and course to course within a content area.
4. Legislation requires the postsecondary-readiness performance standards to be sufficiently rigorous to prepare students in the state to compete academically with students nationally and internationally.
5. Legislation requires a student to meet the cumulative score requirements and achieve a score that meets or exceeds the postsecondary-readiness performance standard on STAAR Algebra II and English III to graduate on the Distinguished Achievement Program.
6. Legislation requires a student to meet the cumulative score requirements and achieve a score that meets or exceeds the passing performance standard on STAAR Algebra II and English III to graduate on the Recommended High School Program.
7. A student may not receive a high school diploma until he or she has met assessment requirements on the STAAR end-of-course assessments.
8. The modified and alternate assessments may have different performance labels than the general assessments have.
9. The general, modified, and alternate assessments should have different key words and phrases in the policy definitions.

Panelists worked in small groups to brainstorm key words and phrases to be used in developing the policy definitions for the three performance levels. To foster a common understanding of the performance levels as the labels were being discussed, groups used "placeholder" labels—Level III, Level II, and Level I—with Level III being the highest level of performance and Level I being the lowest. For purposes of discussion, Level II was considered "passing." The committee was asked to consider the range of student performance within each category but to focus on the students in the middle of the category when making recommendations for key words and phrases to be used in drafting the policy definitions. The committee was also reminded that the TEC requires Level III performance on STAAR Algebra II and English III to indicate postsecondary readiness.

STEP 2: SHARE RECOMMENDATIONS FOR KEY WORDS AND PHRASES

Once the small groups completed the brainstorming activity, a panelist representing each group was asked to share major points from the group's discussion and the group's recommendations for key words and phrases.

STEP 3: REACH CONSENSUS ON RECOMMENDATIONS FOR KEY WORDS AND PHRASES TO BE USED IN DEVELOPING THE POLICY DEFINITIONS

After the small groups shared their recommendations, each group was asked to present key words and phrases for the committee's consideration. As the committee worked toward reaching consensus on recommendations for key words and phrases, the following general comments from the group were captured.

Level III

- As outlined by legislation, Level III should represent postsecondary readiness for STAAR Algebra II and English III in that students performing at this level have the tools and academic preparation needed to be successful in college or a career. The committee preferred to use the phrase postsecondary readiness rather than college and career readiness.
- Performance at this level indicates a high probability of success at the next level without intervention.
- Students who perform at this level demonstrate a deep understanding and insightful application of content. They demonstrate higher-order thinking skills—perhaps the synthesis and evaluation levels of Bloom's taxonomy.
- Students who perform at this level are independent learners and do not need support to make academic progress.

Level II

- Students who perform at this level should be prepared for a variety of postsecondary options (a two-year or four-year degree, a certificate program, or a career). Students entering the workforce need the same set of skills college-bound students need.

- Performance at this level indicates that a student is on track and prepared to be successful at the next level with support.
- Level II may represent a wide range of student performance. Because Level I describes only low-level performance and Level III only high-level performance, there may be a broad range of student performance within Level II, making it difficult to define students in the middle of the category without considering students at both ends of the Level II range (lower end and upper end). The committee suggested dividing Level II into two performance subcategories.
- Students at the upper end of Level II should be successful in entry-level college courses after completing no more than two years of developmental education.

Level I

- Level I should provide a warning sign to students, parents, teachers, and district staff.
- The definition for Level I should communicate a sense of urgency and a substantial need for intervention.
- Use of the word "failing" in the definition was considered. However, the committee did

not want students to be labeled as “failures”; instead, they wanted to communicate in a way that would motivate students to improve.
- The committee wanted to avoid any language in the definitions that implied that students in this category did not have the capacity to achieve academically, especially since the test is a one-day measure of student performance.

At the end of the group discussion on key words and phrases, the following key concepts emerged at all performance levels:

- Level of support or intervention required
- Degree of understanding demonstrated/ability to apply content and skills
- Prediction or likelihood of success at the next level

Identification of these key concepts helped the committee reach consensus on a recommendation of the key words and phrases to be used in developing the STAAR policy definitions.

After the discussion of the STAAR policy definitions for students in general education, the committee was asked to think about the modified and alternate assessments for students receiving special education services and to provide recommendations for issues that TEA should consider in adapting the policy definitions from the general assessments. The following ideas were generated.

- Add “modifications” to the definitions for the modified assessments
- Include links to the academic content for the alternate assessment. The links are identified in the individualized education program (IEP) by the admission, review, and dismissal (ARD) committee.
- Consider noting the relationship to the minimum graduation plan in the policy definition, since students receiving modified or alternate instruction will be graduating on this plan
- Avoid negative connotations or focusing on weaknesses in the descriptions

STEP 4: BRAINSTORM PERFORMANCE LABELS FOR EACH OF THE PERFORMANCE CATEGORIES

The following general guidelines for developing the performance labels were shared.

- The performance labels must clearly represent student performance in each performance category.
- The performance labels must differentiate across the three levels of achievement.
- The performance labels must avoid unnecessary positive or negative interpretations of students themselves.

In their small groups, panelists were asked to brainstorm labels for three levels of performance. The groups were also asked to brainstorm labels for four levels of performance to address

concerns raised about defining the wide range of students within Level II of a three-category system.

STEP 5: SHARE RECOMMENDATIONS FOR PERFORMANCE LABELS

Once the small groups completed the brainstorming activity, a representative from each group was asked to share major points from the group's discussion as well as its recommendations for performance labels.

STEP 6: REACH CONSENSUS ON RECOMMENDATION FOR PERFORMANCE LABELS

After each group shared its recommendations, the committee was led through a discussion to reach consensus on recommendations for the performance labels. The following ideas were generated in the discussion about the labels for three levels of performance:

- STAAR is an assessment of student achievement, so it may make sense to include the word achievement in the labels.
- The label for Level II should represent the wide range of student performance.
- It is important to avoid communicating that a student has "met" the standard for Level II because it is difficult to motivate the student to do better if he or she has already "met" the passing requirement. Panelists also noted that the term "met standard" is too similar to TAKS.
- Although the labels should not be unnecessarily negative, the committee wanted Level I to indicate that something needs to be done to help students performing at this level.
- It might be appropriate to tie the labels to the name of the program—State of Texas Assessments of Academic Readiness—by using the phrase "academic readiness" in the labels.
- The committee also thought it may be possible to use a three-category system and indicate in reporting and communication that a student's performance is at the lower end of Level II rather than subdividing one of the performance levels (Level I or Level II).

The groups then discussed labels for a potential four-category system. The following ideas were shared:

- There was consideration of whether the split should be made to Level I (not passing) or to Level II (passing—middle level). The committee was asked to focus on creating four hierarchical labels that would be used regardless of whether the split subdivided Level I or Level II.
- The committee recommended avoiding the word "approaching" in a passing category.
- The committee generally liked "advanced" for the top category and "insufficient" for the bottom category. In creating a four-level system, panelists wanted to find a word that was more positive than "adequate" for the higher level and less positive than "adequate" for the lower level.

After this discussion, the committee made its recommendations for the performance labels.

Outcome of the PDAC

The committee made the following recommendations for key words and phrases to be used in developing the policy definitions.

Level III

- Postsecondary, college and career ready
- Strongly prepared for success at the next level
- High probability of success at the next level (without intervention or remediation)
- Advanced, deep understanding of knowledge and skills covered by the content standards
- Insightful application of grade-level knowledge and skills
- Demonstrate critical-thinking skills in diverse contexts at an advanced level
- Thoroughly able to manage/manipulate information within a given context
- Independent

Level II

- Adequate, on pace, or prepared for success at the next level, with a possible need for support or targeted interventions
- For students at the upper end of Level II, demonstrate acceptable progress and understanding of content standards, proficient in grade-level knowledge and skills with minimal interventions that may be necessary for success at the next grade level or postsecondary
- For students at the lower end of Level II, partial mastery of grade-level knowledge and skills, fundamental/basic/essential

Level I

- Inadequately prepared for the next level
- Lacking some fundamental knowledge and skills
- Does not demonstrate grade-level knowledge and skills
- Substantial, urgent interventions necessary
- Some knowledge and comprehension but not at the level required to successfully progress
- Serious likelihood of failure at the next level without substantial and immediate intervention

If three performance levels were used for STAAR, the committee had the following recommendations for performance labels, listed in order of preference:

Recommendation 1

Advanced Academic Readiness
Adequate Academic Readiness
Insufficient Academic Readiness

Recommendation 2
- Advanced Achievement
- Adequate Achievement
- Insufficient Achievement

Recommendation 3
- Accomplished Achievement
- Sufficient Achievement
- Limited Achievement

The committee made two recommendations for four performance levels and ranked their suggestions as first choice and third choice to clearly indicate that the first choice was preferred. The recommendations for four performance level labels were as follows:

Recommendation 1
- Advanced Academic Readiness
- Proficient/Satisfactory Academic Readiness
- Limited Academic Readiness
- Insufficient Academic Readiness

Recommendation 3
- Advanced Proficiency
- Proficient
- Approaching Proficiency
- Insufficient Proficiency

Following the meeting, TEA staff used the PDAC recommendations to draft final TEA staff recommendations for performance labels and policy definitions. These staff recommendations were presented to a representative group of PDAC members and received their unanimous approval. The Commissioner of Education subsequently approved the recommendations. There would be two cut scores that would identify three performance categories. For the general STAAR assessments, STAAR Spanish, and STAAR L, the labels for the performance categories are:

- Level III: Advanced Academic Performance
- Level II: Satisfactory Academic Performance
- Level I: Unsatisfactory Academic Performance

Below are the policy definitions for the general STAAR, STAAR Spanish, and STAAR L assessments.

LEVEL III: ADVANCED ACADEMIC PERFORMANCE*

Performance in this category indicates that students are well prepared for the next grade or course. They demonstrate the ability to think critically and apply the assessed knowledge and

skills in varied contexts, both familiar and unfamiliar. Students in this category have a high likelihood of success in the next grade or course with little or no academic intervention.

* *For Algebra II and English III, this level of performance also indicates students are well prepared for postsecondary success.*

LEVEL II: SATISFACTORY ACADEMIC PERFORMANCE**

Performance in this category indicates that students are sufficiently prepared for the next grade or course. They generally demonstrate the ability to think critically and apply the assessed knowledge and skills in familiar contexts. Students in this category have a reasonable likelihood of success in the next grade or course but may need short-term, targeted academic intervention.

** *For Algebra II and English III, this level of performance also indicates students are sufficiently prepared for postsecondary success.*

LEVEL I: UNSATISFACTORY ACADEMIC PERFORMANCE

Performance in this category indicates that students are inadequately prepared for the next grade or course. They do not demonstrate a sufficient understanding of the assessed knowledge and skills. Students in this category are unlikely to succeed in the next grade or course without significant, ongoing academic intervention.

STAAR Modified has the same performance labels as the general STAAR assessments but different policy definitions. The STAAR Modified performance labels and policy definitions can be found on the STAAR Modified web page at http://www.tea.state.tx.us/student.assessment/special-ed/staarm/.

For STAAR Alternate assessments, the performance labels are

- Level III: Accomplished Academic Performance
- Level II: Satisfactory Academic Performance
- Level I: Developing Academic Performance

The policy definitions for the STAAR Alternate performance labels can be found on the STAAR Alternate web page at http://www.tea.state.tx.us/student.assessment/special-ed/staaralt/.

Chapter 5: Performance Level Descriptors

This chapter provides details about Step 3 of the nine-step STAAR standard-setting process, which focused on developing grade/course specific performance level descriptors (PLDs). The chapter covers the following topics.

- What Are Performance Level Descriptors?
- Approach to PLD Development
- Meeting Purpose
- Summary of PLD Meeting Attendees and Proceedings
- Review and Approval Process

What Are Performance Level Descriptors?

PLDs are statements that articulate the specific knowledge and skills students typically demonstrate at each performance level of a test given for a specific grade or course. The PLDs developed for STAAR provide a snapshot of students' academic characteristics based on performance on a given STAAR assessment and reflect the breadth and depth of the content, skills, cognitive demand, and performance requirements evident in the curriculum standards, the Texas Essential Knowledge and Skills (TEKS).

As a component of the standard-setting process, PLDs served to anchor training activities and guide committee members by establishing a common understanding of expected performance on each STAAR assessment. PLDs were used as a reference for the policy committee members as they considered the recommended ranges for cut scores. PLDs were also used by STAAR standard-setting committees to help ground committee members in the content standards and guide them as they made their recommendations for the scores needed to achieve Level II and Level III on each STAAR assessment, including STAAR L, the linguistically accommodated version of STAAR. In addition to their use in standard setting, PLDs have been published to serve as a tool for classroom instruction and to help educators interpret student performance on the assessments. PLDs can enhance parents' understanding of their child's academic strengths and weaknesses and can help the community at large better understand state test scores and the level of performance required of students on STAAR. PLDs are also a requirement of the U.S. Department of Education (USDE) in their review and approval of state assessments.

Approach to PLD Development

Because STAAR represents an aligned system of assessments, PLDs for the STAAR EOC assessments were established first, with lower grades following once high school performance standards were established. TEA, in conjunction with the THECB, convened committees composed of K–12 and postsecondary educators with specific content knowledge and teaching experience to develop PLDs in reading, writing, mathematics, science, and social studies. Educators with experience teaching English language learners (ELLs) and students served by special education were included on the PLD committees. PLDs for the modified and alternate

assessments were developed on a separate schedule by committees composed of both educators with specific content knowledge and educators with special education expertise who understood the learning progression for students taking those assessments.

The approach used to develop the STAAR PLDs aligns with best practices evident in PLD development literature (Mills & Jaeger, 1998; Loomis & Bourque, 2001; Bejar, Braun & Tannenbaum, 2006; Perie, 2007; Perie, Hess, & Gong 2008). In developing the methodology, TEA consulted a national expert in PLD development (Redfield & Sheinker, 2006). The primary characteristics of the approach are listed below.

- PLDs were developed in advance of standard setting.
- Content experts (K–12 and postsecondary educators, including those with special education and ELL expertise) developed the PLDs.
- Committees were carefully selected for their content knowledge and experience in teaching the content of the test. The size of the committees (seven to ten educators per committee) facilitated in-depth discussion.
- The committee composition was designed to be representative of all students taking the assessments.
- Guidance to the committees including the following characteristics of PLDs:
    - PLDs must connect directly to the knowledge and skills evident in the content standards.
    - PLDs should reflect the range of cognitive demand represented by the content standards.
    - PLDs should describe performance in the middle of the performance category.
    - PLDs should reflect the learning progression evident in the content standards.
    - PLDs apply to all students taking the assessment.
    - PLDs reflect student performance (as opposed to student attitudes toward the content or the test).
    - PLDs describe student performance in relation to the content standards, not specific questions or tasks.
- PLDs underwent a vertical articulation to ensure that their organization reflected the progress in learning across grade levels and courses.
- As part of the standard-setting process, the PLDs were revised to reflect the input of the standard-setting committees and to reflect appropriate inferences based on where performance standards were set.
- PLDs were finalized and made public following the standard-setting process.
- PLDs will be reviewed periodically to maintain alignment as curriculum and/or performance standards are revised.

Meeting Purpose

The purpose of the PLD meetings was to convene Texas K–12 and postsecondary experts to define student achievement within each performance category for all STAAR assessments. The committees were charged with:

1. considering the performance labels and policy definitions, the assessed curriculum, and the culminating skills for each grade/course;
2. developing draft performance level descriptors for Level II: Satisfactory Academic Performance, Level III: Advanced Academic Performance, and Level I: Unsatisfactory Academic Performance for each grade/course;
3. reviewing the performance level descriptors for all three levels and adjusting as necessary to reflect student performance across the performance categories;
4. reaching consensus on the recommendations for the performance level descriptors for each grade/course; and
5. reviewing the performance level descriptors across grades/courses within a content area for reasonableness.

Summary of PLD Meeting Attendees and Proceedings

The STAAR EOC PLD meetings convened in November 2011. The social studies and science committees met on November 2–3, and the English and mathematics committees met on November 8–9. The STAAR 3–8 PLD meetings convened in June 2012. The writing, grades 6–8 reading, and grades 6–8 mathematics committees met on June 11–12; and the social studies, science, grades 3–5 reading, and grades 3–5 mathematics committees met on June 21–22.

COMMITTEE COMPOSITION

The PLD committees were composed of seven to ten K–12 and postsecondary educators, including those with special education and ELL experience. Tables 5.1 and 5.2 summarize the PLD committee composition for the STAAR EOC and 3–8 PLD committees, respectively.

Table 5.1: STAAR EOC PLD Committee Composition

Gender		Ethnicity		Position		Classroom Experience with Student Population	
Male	30	Native American	2	Teacher, General	53	Special Education	60
Female	52	Asian/Pacific Islander	1	Teacher, Special Education	3	ELL	51
		African American	10	Teacher, ESL/Bilingual	7		
		Hispanic	20	Other Assignment*	9		
		White	49	Higher Education	10		

*Other assignment includes curriculum coordinator, curriculum manager, specialist, and department head.

Table 5.2: STAAR 3–8 PLD Committee Composition

Gender		Ethnicity		Position		Classroom Experience with Student Population	
Male	22	Native American	0	Teacher, General	76	Special Education	88
Female	91	Asian/Pacific Islander	1	Teacher, Special Education	2	ELL	84
		African American	15	Teacher, ESL/Bilingual	3		
		Hispanic	32	Other Assignment*	30		
		White	65	Higher Education	2		

*Other assignment includes curriculum coordinator, specialist, facilitator, director, principal, and department head.

MEETING PROCEEDINGS

Before convening the PLD committees, TEA content experts met to consider the assessed curriculum standards with the goal of identifying a preliminary set of culminating skills. This exercise was critical for several reasons: 1) it served as a training tool for the PLD meeting facilitators to begin thinking about the curriculum in terms of a performance continuum; 2) it provided a framework for articulating the "big ideas" in the content standards; and 3) the culminating skills served as a starting point for the committees in thinking about how to organize the content into PLDs.

PLD committees for STAAR were convened for two-day meetings. On Day 1, the committees met jointly for a program overview and an orientation to the task of developing PLDs. The orientation included the following information:

- an overview of the goals and organization of the new program, including the focus on readiness standards, the goals for rigor in the performance standards and assessments, and the aligned nature of the STAAR assessments;
- an introduction to the graduation requirements by diploma plan and the ways in which the requirements relate to achievement at the different performance levels;
- an overview of the standard-setting process, including a discussion on how the PLDs would be used as input to the policy committee and the standard-setting committees as they made recommendations for cut-scores;
- an introduction to the performance labels and policy definitions; and
- a general orientation to PLDs, as well as specific orientation to the committee tasks.

Following group orientation, the panelists separated into their respective committees to develop PLDs. Committee discussion was facilitated by TEA content experts trained in PLD development.

In the process of developing PLDs, committee members considered the performance labels and policy definitions, the assessed curriculum, and the draft culminating skills for each assessment. The committees began with a discussion of what the performance labels and policy definitions indicate about student performance in relation to the content being assessed. Each committee then considered the content associated with Level II: Satisfactory Academic Performance.

Committee members reviewed and discussed the culminating skills, revised those skills as necessary to reflect satisfactory performance, and organized the information into a bulleted list of satisfactory-level PLDs.

Once the committees reached consensus on the PLDs for satisfactory-level performance, they moved up to Level III: Advanced Academic Performance, considering what extended skills from the curriculum students would need in order to demonstrate advanced proficiency. They also moved down to Level I: Unsatisfactory Academic Performance and identified the prerequisite or enabling skills from the curriculum that students at that level could demonstrate. Figure 5.1 illustrates this process.

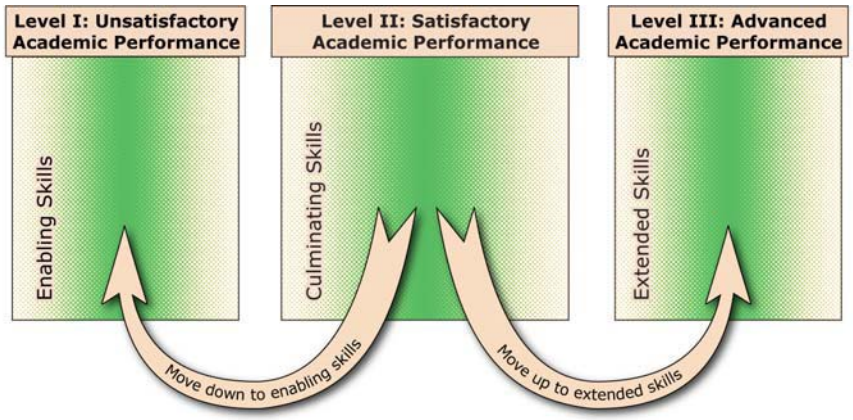


Figure 5.1: PLD Development Process

During this process, committees were advised to avoid the use of adjectives and adverbs to describe the level of proficiency (e.g., less, seldom, sometimes, partially, rarely, minimal) and to focus on the application of content and skills from the curriculum that differentiated performance at each level.

After drafting the PLDs for a particular grade/course, each committee reviewed the PLDs to ensure that performance moves from the lowest to highest across the three levels. The following guidelines were used.

- Do levels of cognitive demand increase?
- Do the PLDs represent the range of student performance expected for the grade/course as reflected in the curriculum?
- Are mastered skills subsumed and new skills evident when moving through the performance continuum?
- Is repetition eliminated?

On the second day of each meeting, the committees met together by content area (reading, writing, mathematics, science, and social studies) to ensure that the PLDs were well articulated across the grades/courses in a content area. Guiding questions for the combined committees included the following.

- Are levels of cognitive demand represented in each performance category generally parallel?
- Do the PLDs represent the range of expected performance for each grade/course?

At the close of the meeting, a short survey was conducted to solicit information about how well the committee members understood their task and to gather a confidence measure for the PLD process. A summary of the survey responses is provided in Appendix 5. In general, committee members indicated that they agreed their role was clear, the task was well-defined, they felt comfortable expressing their opinion, the time provided was sufficient for the task, and they would be willing to participate in similar activities in the future.

Review and Approval Process for PLDs

Following the PLD committee meetings, TEA staff reviewed the PLDs, applying consistency in formatting and verifying that the articulated content and skills for each level matched the performance expectations evident in the policy definitions. In addition, the PLDs were reviewed by Dr. Jan Sheinker, a nationally recognized alignment expert with broad experience working with the USDE during peer review. Dr. Sheinker provided feedback related to the level of alignment between the PLDs and the content standards. This feedback was incorporated into the final version of the PLDs.

PLD feedback from the standard-setting committees primarily reflected clarifications that the committees found useful in their discussions about student performance during the standard-setting process. Following the standard-setting meetings, this feedback was incorporated into the final version of the PLDs. The final PLDs can be found under Performance Level Descriptors at the STAAR resources page on the TEA website: http://www.tea.state.tx.us/student.assessment/staar/.

Chapter 6: Policy Committee and Neighborhood Development

This chapter provides details about Step 4 of the nine-step STAAR standard-setting process, which focuses on convening the policy committee for EOC assessments (see Table 2.1) and developing performance standard ranges ("neighborhoods") for EOC assessments and grades 3–8 assessments (see Table 2.2). The sections in this chapter include

- Purpose of Neighborhoods
- Purpose and Format of the Policy Committee
- Policy Committee Composition
- Policy Committee Meeting Proceedings
- STAAR EOC Empirical Studies Reviewed by Committee
- STAAR EOC Operational Definitions of Postsecondary Readiness
- STAAR EOC Neighborhood Development Guidelines
- STAAR EOC Neighborhood Recommendations and Rationale
- Policy Committee Surveys
- STAAR 3–8 Empirical Studies
- STAAR 3–8 Neighborhood Development

Purpose of Neighborhoods

Texas Education Code (TEC) §39.0241 requires that performance standards be aligned from grade 3 through end-of-course assessments. Under an aligned set of standards, student performance at each level (i.e., Unsatisfactory, Satisfactory, or Advanced Academic Performance) within a content area should indicate whether or not the student is on track to be successful in the next grade or course. TEA conducted extensive research to support an evidence-based standard setting approach to fulfill the legislative intent. These studies established links between performance on STAAR and performance on other assessments and provided research-based anchors for setting meaningful and rigorous performance standards.

The results of the empirical studies were used to create reasonable ranges or "neighborhoods" in which the performance standards could be set for STAAR assessments. The neighborhoods reflected the results of the empirical studies but not the technical aspects of the various studies. Guidelines for neighborhood development provided a consistent approach to defining the ranges. The neighborhoods were used by the standard-setting committees to recommend performance standards.

A policy committee—composed of policy experts, legislative staff, business and workplace leaders, and secondary- and higher-education representatives—used the study results to inform its recommendations for STAAR EOC neighborhoods. The recommended neighborhoods were used by the standard-setting committees to recommend performance standards for STAAR EOC. The STAAR 3–8 neighborhoods were determined by considering the alignment of performance standards with EOC assessments and by using the results of various studies. Standard-setting committees used the STAAR 3–8 neighborhoods to make performance-

standard recommendations. It was not necessary to convene a policy committee for STAAR 3–8 since the goal was not to establish new policy inferences but to carry the inferences down from STAAR EOC to STAAR 3–8.

Purpose and Format of the Policy Committee

The purpose for convening the policy committee was to obtain recommendations on the reasonable ranges, or neighborhoods, for each performance standard on the STAAR EOC assessments. Committee members reviewed the test purposes, uses of the performance standards, and general definitions of the performance levels. They were presented with the results from validity and linking studies as well as the draft performance level descriptors (PLDs). Using this information and drawing on their policy expertise, the committee was able to provide input about ranges for the STAAR EOC cut scores that would support meaningful inferences about educational outcomes.

COMMITTEE CHARGE

TEA and THECB officially charged the policy committee with the following:

> "The policy committee for the STAAR EOC assessments will recommend reasonable ranges (for use by the standard-setting committee) within which to set the STAAR EOC performance standards by
>
> - providing guidance to TEA and THECB on how to appropriately evaluate the results of the standard-setting research studies; and
> - considering the policy implications of the performance standards so that STAAR EOC cut scores support meaningful inferences about educational outcomes (a student's postsecondary readiness or readiness for the next course)."

MEETING FORMAT

The policy committee meeting took place over one and one-half days (February 1–2, 2012) and consisted of committee members representing diverse stakeholder groups (committee composition is described in the next section). Before the policy committee was convened, all validity and linking studies were completed, summarized, and reviewed by TEA. Committees that developed the STAAR performance labels, policy definitions, and specific performance level descriptors had also already been convened, and their recommendations were presented to the policy committee.

The policy committee meeting was led by two external facilitators, Dr. Gregory Cizek from the University of North Carolina and Dr. Wayne Camara from the College Board. Both facilitators are experts in standard setting. Dr. Cizek was also the facilitator for the Performance Descriptor Advisory Committee (PDAC), which developed the STAAR performance labels and policy definitions, described in Chapter 4 of this report.

The policy committee's neighborhood recommendations were reviewed and incorporated into the materials for the standard-setting committee meetings, which took place in late-February 2012, three weeks following the policy committee meeting.

Policy Committee Composition

The policy committee was composed of 28 members who were educators and administrators at the secondary- and higher-education level, business and workplace leaders, policy experts, legislative staff, and special population representatives from across the state of Texas. Table 6.1 shows the groups from which policy committee members were recruited and the rationale for including each group.

Table 6.1: Groups for Recruiting Policy Committee Members

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 includes 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.

For the purpose of continuity in the STAAR EOC standard-setting process, the policy committee consisted of several members who attended the Performance Descriptor Advisory Committee (PDAC). In addition, policy committee members were invited to observe the standard-setting committees in late-February to hear the discussions and ideas shared during the standard-setting meetings.

Tables 6.2–6.5 summarize the characteristics and experience of the 28 policy committee members. Refer to Appendix 6 for a complete list of the names, positions, and affiliations of the policy committee members.

Table 6.2: Gender Distribution of Policy Committee

Gender	Number	Percentage (out of 28)
Female	12	43%
Male	16	57%

Table 6.3: Ethnicity Distribution of Policy Committee

Ethnicity	Number	Percentage (out of 28)
African American	2	7%
Hispanic	4	14%
White	22	79%

Table 6.4: Current Position and Years of Experience in Education of Policy Committee Panelists

		Years of Professional Experience in Education						
		None	1–5 years	6–10 years	11–15 years	16–20 years	More Than 20 years	Total
Current Position	Business or Workplace	0	1	0	0	0	3	**4**
	Educational Administration	0	0	0	0	2	10	**12**
	Higher Education	0	0	0	1	0	3	**4**
	Legislative	1	0	2	2	2	0	**7**
	Teacher	0	0	1	0	0	0	**1**
	Total	**1**	**1**	**3**	**3**	**4**	**16**	**28**

Table 6.5: Policy Committee's Experience with Student Populations

Student Population	Number	Percentage (out of 28)
General Education	24	86%
Special Education	23	82%
English Language Learners (ELL)	20	71%
Low Socioeconomic Status	24	86%

Policy Committee Meeting Proceedings

During the meeting, the policy committee considered and discussed the policy implications of the STAAR EOC performance standards, including the postsecondary-readiness standards. The committee also considered

- the connection between the standards (Level II: Satisfactory Academic Performance and Level III: Advanced Academic Performance) within a test,

- the connection between the standards across tests within a content area, and
- the connection between the standards across content areas.

All committee discussions were informed by the empirical study results. Table 6.6 shows the agenda for the policy committee meeting.

Table 6.6: Policy Committee Meeting Agenda

Day 1	• Welcome and Introductions • Background and Overview of Policy Committee Meeting • Discussion of Policy Questions • Part 1: Committee Judgment and Feedback
Day 2	• Part 2: Committee Judgment and Feedback • Part 3: Committee Judgment and Feedback • Cross-Content Area Articulation and Final Recommendations • Evaluation and Closing Remarks

A description of each topic in the agenda is detailed below.

WELCOME AND INTRODUCTIONS

The committee members were introduced and general housekeeping tasks were discussed, including the non-disclosure agreement, security protocols, and reimbursement forms.

BACKGROUND AND OVERVIEW OF POLICY COMMITTEE MEETING

TEA and Pearson staff provided background information about the STAAR program and the EOC assessments, federal and state legislative requirements, performance categories and policy definitions for STAAR, and an overview of the evidence-based standard-setting process for STAAR EOC, including the role of the policy committee.

DISCUSSION OF POLICY QUESTIONS

Committee members were asked to provide their own answers to a set of policy questions as a way of communicating their expectations for student success on the STAAR assessment overall and in relation to other assessments for which data were collected. The following policy questions were presented to the committee:

- How should the Level II and Level III standards for STAAR Algebra II and English III compare to
  - college admissions tests?
  - college placement tests?
  - TAKS?
  - NAEP?

- In general, what percentage of students would you expect to be in each performance category (Level I, II, and III)?
- What type of consistency in passing rates is expected among STAAR EOC assessments across content areas (e.g., English, mathematics, science, and social studies)?

The purpose of these policy questions was to help the committee members think about their preconceived expectations for the new assessment program and its relation to other criteria or measures. The committee members reviewed their answers to these questions throughout the meeting after they examined the empirical study results and participated in rounds of discussion.

COMMITTEE JUDGMENT AND FEEDBACK

Committee members were organized into five table groups. Within each group, members went through a process in which they examined the results of the empirical studies, discussed these results in relation to the policy questions they had previously answered, provided judgments or recommendations about neighborhoods for the cut scores (Level II and Level III) on each STAAR assessment, and gave summary feedback about the judgments.

This process was done in three parts. In Part 1, the committee focused on the studies and neighborhoods for STAAR Algebra II, English III reading, and English III writing, the three assessments for which indicators of postsecondary readiness were required by statute. In Part 2, the committee considered the remaining mathematics and English STAAR EOC assessments. In Part 3, the committee focused on the science and social studies STAAR EOC assessments.

The committee was provided information about the data and designs used for each empirical study, the history and purpose of the different assessments or measures in each study, and the implications and limitations of each study. Because there were a large number of empirical studies for the committee to review in a relatively short amount of time (1.5 days), it would have been overwhelming to present the studies one at a time and ask committee members to make recommendations about each study. To make the process more manageable, TEA and THECB constructed “neighborhood options” as starting points for the committee to consider. Each neighborhood option represented certain assumptions about the empirical studies that yielded particular neighborhood bounds for each cut score. More details about how the neighborhood options were constructed are provided in the section “Neighborhood Development Guidelines.”

After reviewing the studies and participating in the table- and group-level discussions, committee members were asked to provide their individual judgments by rank-ordering the neighborhood options. Judgments were collected for each content area (mathematics, English reading, English writing, science, and social studies) during the three-part process.

CROSS-CONTENT AREA ARTICULATION AND FINAL RECOMMENDATIONS

After committee members had given their judgments and were provided feedback for all content areas, they were asked to consider the recommended neighborhood options as a whole. Specifically, the committee was asked these questions:

- Do the recommended options make sense as a comprehensive assessment system?
- What changes or "tweaks" does the committee recommend for the neighborhoods?
- What other recommendations does the committee have about the proposed STAAR EOC performance standards?

After discussing these questions, the committee made its final recommendations about the neighborhoods and provided rationales for its recommendation.

EVALUATIONS AND CLOSING REMARKS

TEA and THECB thanked committee members for their work over the 1.5-day meeting. They provided an overview of how the committee's recommendations would be used in the remaining steps of the STAAR EOC standard-setting process along with the timeline for each step. Committee members also filled out process evaluation surveys.

STAAR EOC Empirical Studies Reviewed by Committee

The empirical studies were a key component in helping inform the policy committee's discussions and recommendations. Results from all the validity and linking studies conducted were presented at various points and in different formats during the policy committee meeting. Table 6.7 lists the studies that were presented to the committee during each judgment and feedback part of the meeting.

Table 6.7: Validity and Linking Studies Presented to Policy Committee

Part of Meeting	STAAR Assessments	Empirical Studies
Part 1: Judgment and Feedback	STAAR Algebra II STAAR English III reading STAAR English III writing	• External validity studies o Comparisons with SAT and ACT o Comparisons with THEA and ACCUPLACER o College Students taking STAAR o Comparisons with NAEP • STAAR–TAKS comparison studies • Grade correlation studies
Part 2: Judgment and Feedback	STAAR EOC mathematics STAAR English reading STAAR English writing	• STAAR–STAAR linking studies

Table 6.7 cont.: Validity and Linking Studies Presented to Policy Committee

Part of Meeting	STAAR Assessments	Empirical Studies
Part 3: Judgment and Feedback	STAAR EOC science STAAR EOC social studies	• External validity studies o Comparisons with SAT and ACT o Comparisons with NAEP • STAAR–TAKS comparison studies • Grade correlation studies

The main method for presenting empirical study results was to use an **empirical number line**, such as the one shown in Figure 6.1. Empirical number lines were integrated into the facilitators' PowerPoint presentation. Committee members also received paper copies of the empirical number lines in their binders. Figure 6.1 shows the empirical number line used for STAAR Algebra II. The full set of empirical number lines used in the policy committee is given Appendix 7.

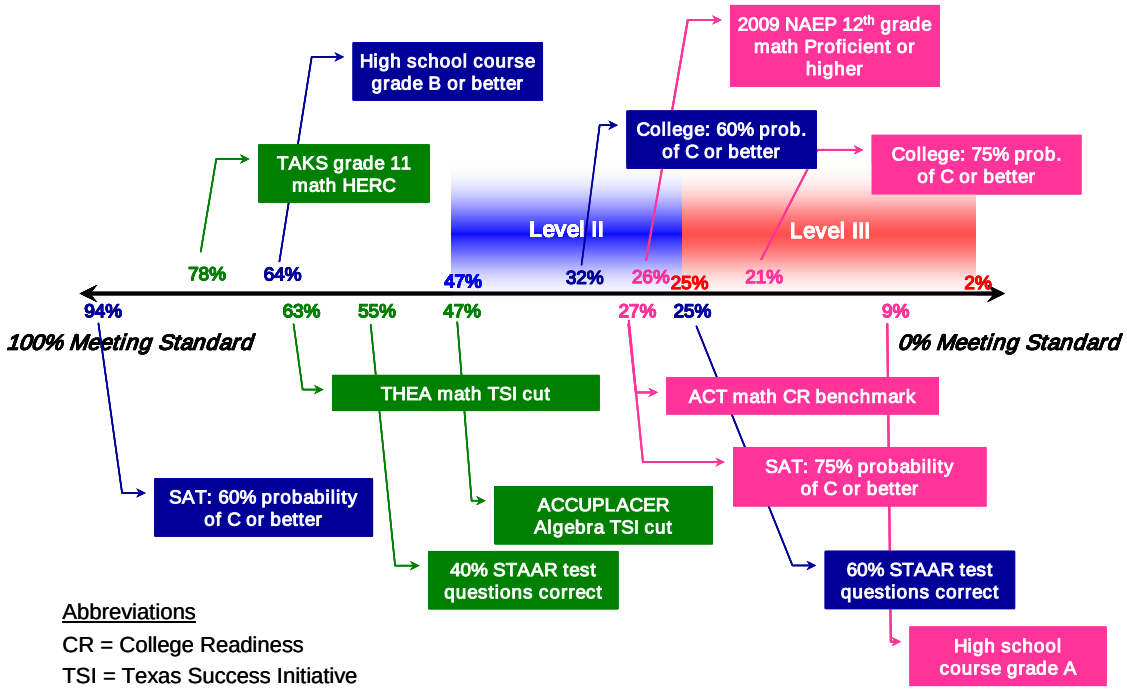


Figure 6.1: Empirical Number Line for STAAR Algebra II

In addition, the validity studies **quality summary** and full set of **study profiles** were available in binders located on the resource tables in the committee meeting room. Committee members were invited to review these materials and ask any questions that arose about them throughout the course of the meeting.

Refer to Chapter 3 for more detailed descriptions of the empirical number lines, quality summary, and study profiles.

Operational Definitions of Postsecondary Readiness

In order for the policy committee to use the empirical studies in the context of the STAAR EOC performance standards, it was important to first present operational definitions for the performance standards to connect them to the study results, especially as they relate to measuring postsecondary readiness, one of the key goals of the STAAR program.

Using the definition of college readiness provided in statute (TEC, Section 39.024; see Appendix 1) and the general policy definition recommended by the PDAC (see Chapter 4), the following specific operational definitions of postsecondary readiness were crafted for the Level II and Level III performance standards on STAAR Algebra II and English III.

Level II: Satisfactory Academic Performance Operational Definition

- Students in this category are reasonably likely (with at least a 60% probability) to succeed (with a grade of C or higher) in an entry-level, credit-bearing course in that content area for a baccalaureate degree or associate degree program at a general academic teaching institution or a postsecondary institution that primarily offers associate degrees, certificates, or other credentials.

Level III: Advanced Academic Performance Operational Definition

- Students in this category are highly likely (with at least a 75% probability) to succeed (with a grade of C or higher) in an entry-level, credit-bearing course in that content area for a baccalaureate degree or associate degree program at a general academic teaching institution or a postsecondary institution that primarily offers associate degrees, certificates, or other credentials.

These operational definitions were presented to the policy committee before Part 1 of the committee judgment and feedback. Committee members were given time to discuss the operational definitions. This allowed committee members to think about and internalize these definitions before the committee was presented with any empirical study results for STAAR Algebra II and English III.

STAAR EOC Neighborhood Development Guidelines

"Neighborhoods" are reasonable ranges in which the performance standards may be set for each STAAR assessment. The main charge of the policy committee was to make neighborhood recommendations within which cut scores for Levels II and III could be set. The standard-setting committees to follow would then make cut score recommendations within the neighborhoods. Figure 6.2 illustrates the relationship between neighborhoods and cut scores. It also contrasts the roles of the policy committee and the standard-setting committees.

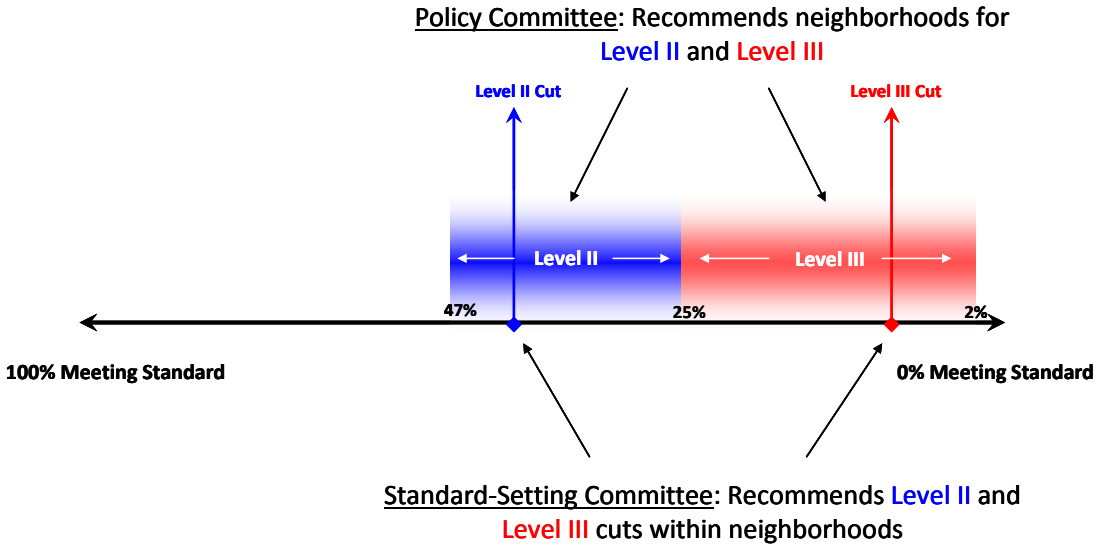


Figure 6.2: Relationship of Neighborhoods and Cut Scores

TEA and THECB constructed guidelines for developing the neighborhoods for each STAAR assessment. The guidelines were grounded in the operational definitions for Level II and III and informed by the list of available empirical studies and measures for each assessment. Figure 6.3 illustrates the process of developing neighborhoods.

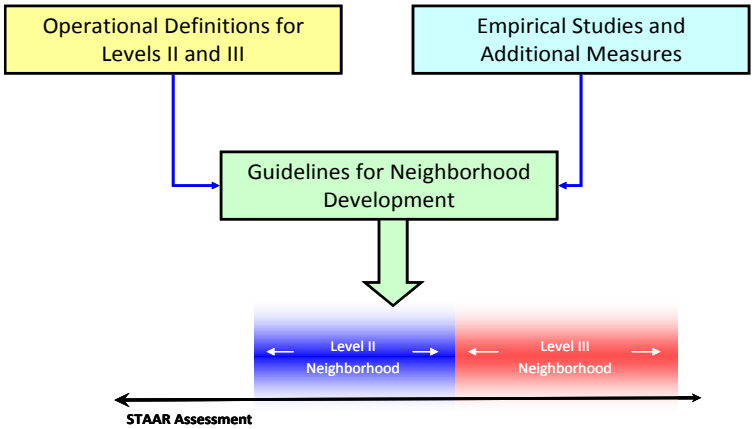


Figure 6.3: Neighborhood Development Process

Figures 6.4 and 6.5 show the neighborhood development guidelines for STAAR EOC assessments. Figure 6.4 provides the details of the guidelines for each performance standard. Figure 6.5 illustrates the guidelines graphically. Both were shared with the policy committee.

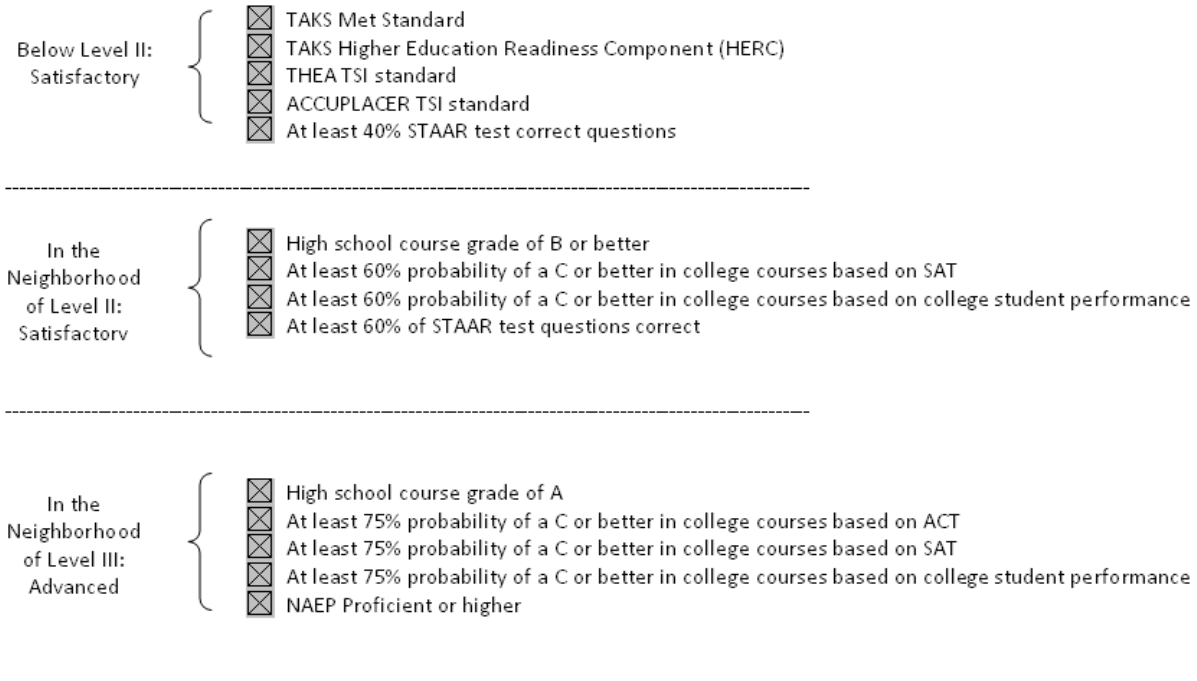


Figure 6.4: STAAR EOC Neighborhood Development Guidelines

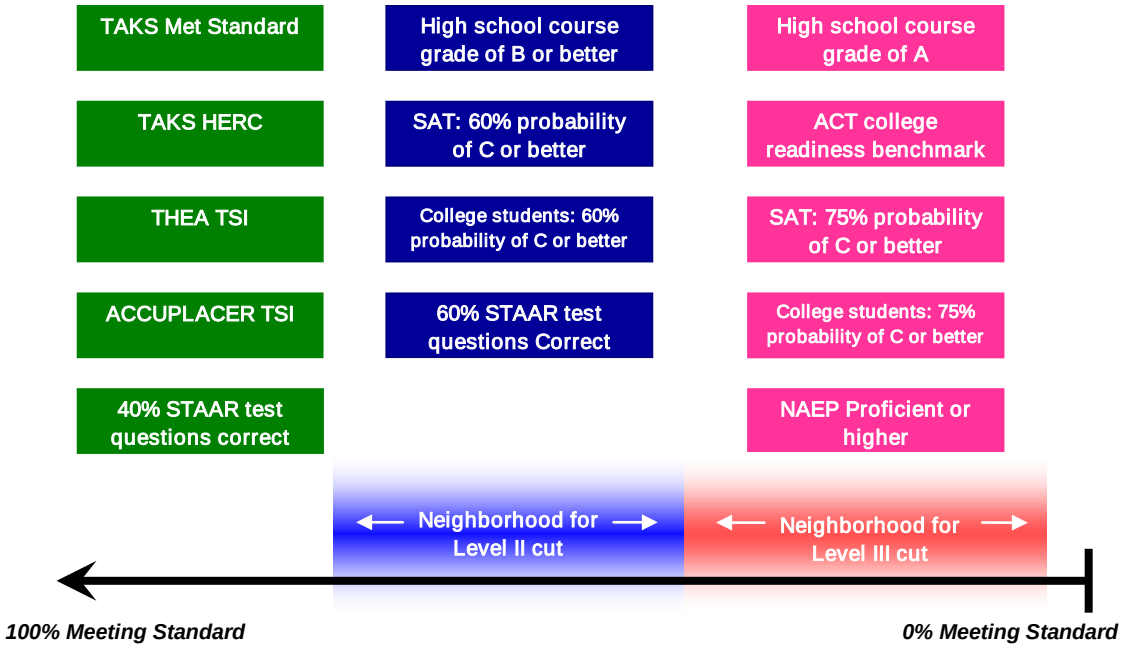


Figure 6.5: Graphical Illustration of STAAR EOC Neighborhood Development Guidelines

Several guidelines shown in Figures 6.4 and 6.5 refer to the Texas Success Initiative (TSI) standards. TSI is a requirement in the TEC (Section 51.3062) that calls for incoming college freshmen to be assessed in reading, writing, and mathematics before their enrollment in an institution of higher education in Texas. TSI standards are cut scores established on external

assessments authorized by THECB. The cut scores indicate whether incoming students need remedial coursework before enrolling in any entry-level college courses in the same content area. In other words, the TSI standards on the external assessments are indicators of readiness for entry-level college coursework. For more information about TSI, refer to the THECB website at: http://www.thecb.state.tx.us/facts/cd/Page4.htm.

NEIGHBORHOOD OPTIONS

As shown in Figures 6.4 and 6.5, a substantial number of empirical studies were conducted that could be used to develop the neighborhoods for the 15 STAAR EOC assessments. The amount of data would have been overwhelming for committee members to process in a 1.5-day meeting, especially given that the purpose of these data were to help the committee make informed recommendations. To make the process more manageable, "neighborhood options" were developed for committee members to consider.

Each neighborhood option represented certain assumptions about the empirical studies that yielded particular neighborhood bounds for each cut score. Three neighborhood options were developed for each of the assessed content areas (mathematics, English reading, English writing, science, and social studies). The assumption underlying each option was as follows:

- Option A: Developed from the operational definitions and empirical study results
- Option B: Option A adjusted for motivation-effect estimates
- Option C: Additional upward adjustment beyond Option B

Option B was developed in response to low student motivation in the data collected for the empirical studies. The impact data (i.e., the percentage of students projected to meet each cut score) shown in the empirical studies number lines are based on student performance during the spring 2011 administrations of the EOC assessments. These assessments were not high stakes for test takers in 2011. Some assessments (English II, English III, and world history) were administered as stand-alone field tests, for which no test scores were reported. It was expected that after the STAAR EOC assessments become the graduation testing requirement, students are likely to be more motivated when testing. Therefore, the percentage of students in each performance level is likely to be greater than what was observed in spring 2011. To develop neighborhoods for Option B, TEA and THECB looked at historical trend data based on TAKS, specifically what changes in passing rates occurred between 2003 and 2004, when TAKS exit level (grade 11) became the graduation testing requirement. Other factors, such as the shift in the difficulty of assessments between TAKS and STAAR, were also considered. It was expected that the difficulty increase between TAKS and STAAR will be higher than the increase seen between TAAS and TAKS; therefore, motivation may not have as big an effect for STAAR as it did for TAKS. Students may be more motivated in spring 2012, but they may not be as prepared for the significant increase in rigor. STAAR assesses nearly all the student expectations for each course rather than a selected subset, which made up the grade-level high school assessments under TAKS. Performance on STAAR is expected to increase over time as instruction is adjusted to meet the new expectations. Table 6.8 shows the motivation adjustment estimates applied to the impact data in Option A for each content area.

Table 6.8: Motivation Adjustment Estimates by Content Area

Content Area	Level II: Satisfactory	Level III: Advanced
Mathematics	15%	5%
English (Reading and Writing)	20%	5%
Science	10%	5%
Social Studies	15%	5%

The rationale for the motivation adjustment values was as follows:

- Adjustments for Level III were less than for Level II. This is consistent with historical trend data in Texas. There are fewer students in the higher performance level to begin with, and high-performing students are typically less affected by motivation.
- For mathematics, English, and science, the motivation effect for STAAR was estimated (and adjusted) to be slightly lower than the motivation effect that was observed for TAKS. For those three areas, there was a significant shift in the content assessed. For example, in mathematics, Algebra II content was not assessed on any of the TAKS tests. On STAAR English assessments, students are required to respond to more genres of reading and to write different types of essays, such as analytical and persuasive. In science, chemistry and physics content is assessed on STAAR, as opposed to the lower-level integrated physics and chemistry content assessed on TAKS.
- For social studies, the motivation effect for STAAR was estimated (and adjusted) to be slightly higher than the motivation effect that was observed for TAKS. This is because the TAKS social studies performance standards were set relatively lower than for the other content areas. It was expected that the social studies standards for STAAR would more closely align with the other content areas, increasing the potential impact of motivation.
- English had the largest motivation adjustment at Level II because of the assessments' open-ended items (short answer and essays). Because performance tasks require a considerable effort from students, these types of items are most likely to be affected by motivation. In addition, performance tasks on both the reading and writing assessments are weighted so that they make up a significant percentage of the overall test score.
- Science had the smallest motivation adjustment. Historically, TAKS data suggest that science has one of the lowest motivation effects. In addition, the science assessments had one of the greatest shifts in content difficulty from TAKS to STAAR.

It should be noted that there is not a substantial amount of data to inform the motivation adjustment estimates. The research literature is inconclusive in this regard; there are many factors that can influence motivation, and it is difficult to isolate the effects of any one factor.

Option C was developed as an additional upward adjustment beyond Option B. Because it represented higher expectations for the performance standards (Level II and Level III), it created a larger estimated motivation effect. For instance, "success" in an entry-level, credit-bearing course could be defined as obtaining a grade of B or better (instead of C or better) in the related content area. This would change the neighborhood development guidelines (shown in Figures 6.4 and 6.5) and lead to a set of neighborhoods whose boundaries would indicate higher standards than in Options A or B.

Figure 6.6 provides, as an example, the neighborhood options developed for STAAR Algebra II.

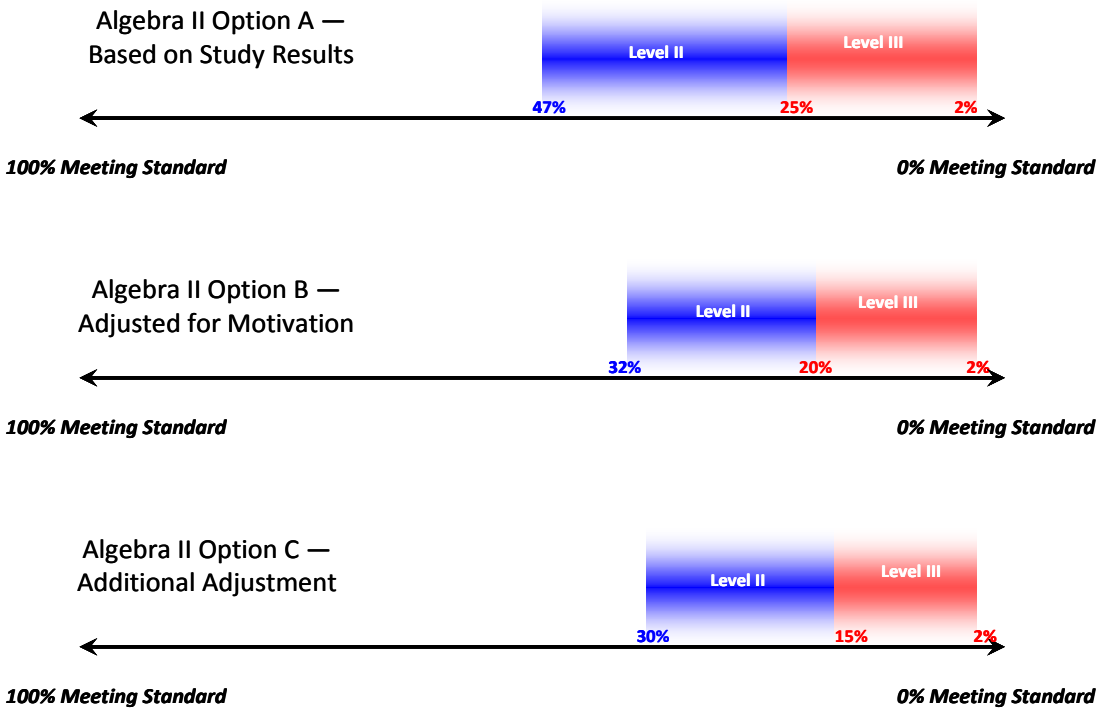


Figure 6.6: Neighborhood Options for STAAR Algebra II

It is important to note the impact on the size of the neighborhoods after the motivation adjustment was applied in Option B and the additional adjustment was applied in Option C. Application of the motivation factor to Option B reduced the neighborhood within which the standard-setting committee could work. The additional adjustment applied in order to obtain Option C further reduced the neighborhood for each performance standard.

The complete set of neighborhood options shown to the policy committee is provided in Appendix 8.

STAAR EOC Neighborhood Recommendations and Rationale

Each committee member was asked to provide judgments by rank-ordering the neighborhood options, with 1 indicating most preferred and 3 representing least preferred, for each content

area during the three-part judgment and feedback portion of the meeting. The judgments were tallied and the neighborhood option receiving the most first-place ranks was designated as the committee's recommendation for each content area.

Tables 6.9–6.11 show summaries of the committee's judgments after each part of the meeting. The **bolded** option for each content area represents the committee's recommendation for that part of the meeting. As indicated in the tables, Option A was most preferred by most committee members for every content area during all parts of the meeting.

Table 6.9: Summary of Committee Judgments (Part 1[3])

STAAR Algebra II

Option	A	B	C
Rank = 1	**18**	8	0
Rank = 2	3	18	5
Rank = 3	5	0	21

STAAR English III Reading

Option	A	B	C
Rank = 1	**19**	7	0
Rank = 2	2	18	6
Rank = 3	5	1	20

STAAR English III Writing

Option	A	B	C
Rank = 1	**17**	9	0
Rank = 2	3	17	6
Rank = 3	6	0	20

Table 6.10: Summary of Committee Judgments (Part 2)

STAAR EOC Mathematics

Option	A	B	C
Rank = 1	**21**	6	0
Rank = 2	2	21	4
Rank = 3	4	0	23

STAAR EOC English Reading

Option	A	B	C
Rank = 1	**20**	7	0
Rank = 2	5	19	3
Rank = 3	2	1	24

STAAR EOC English Writing

Option	A	B	C
Rank = 1	**20**	7	0
Rank = 2	5	20	2
Rank = 3	2	0	25

Table 6.11: Summary of Committee Judgments (Part 3)

STAAR EOC Science

Option	A	B	C
Rank = 1	**21**	6	0
Rank = 2	4	21	2
Rank = 3	2	0	25

STAAR EOC Social Studies

Option	A	B	C
Rank = 1	**17**	9	1
Rank = 2	6	18	3
Rank = 3	4	0	23

After cross-content area articulation, during which the committee looked at the recommended neighborhoods as a comprehensive assessment system, the committee still preferred Option A for all content areas. The final recommendations by the policy committee are summarized in Table 6.12.

Table 6.12: Final Neighborhood Recommendations by Policy Committee

Content Area	Recommended Neighborhood Option
Mathematics	**Option A**
English Reading	**Option A**
English Writing	**Option A**

[3] One panelist left the meeting prior to making any judgments and did not return. One panelist was unavailable near the end of Day 1, when the judgments for Part 1 were collected but returned for Part 2 and Part 3.

Table 6.12 cont.: Final Neighborhood Recommendations by Policy Committee

Content Area	Recommended Neighborhood Option
Science	**Option A**
Social Studies	**Option A**

The committee's main reason for its recommendations centered on the uncertainty of the projected impact data (percentage of students meeting each cut score). Panelists were well aware that the impact data were based on unmotivated student responses. They noted that it was difficult to accurately estimate or predict the motivation effect because of the significant changes in the assessed curricula between TAKS and STAAR, changes in available resources at districts and campuses, and considerations about teacher training and development. Therefore, the general consensus was to recommend the neighborhood options based directly on the empirical studies.

The committee was also given the flexibility of "tweaking" the options as part of its final recommendations. The committee provided the following feedback.

- Some committee members thought the jump from Option A to Option B was too great and would have preferred an option between the two options.
- Some committee members thought the lower bounds of the Level II neighborhoods were too low and would like to raise them by about 5%. Others felt the range (both lower and upper bounds) of the Level II neighborhood should be lowered.
- Many committee members thought the upper bounds of the Level III neighborhoods were too high in general and recommended that they be lower.
- Several committee members thought the range for the Level III neighborhoods was too large, especially for STAAR U.S. history and world geography. They suggested the lower bound of the Level III neighborhoods be raised.

Following the policy committee meeting, TEA considered each piece of feedback. However, because there was no consensus among committee members on these points and several of the points were in opposition to one another, TEA decided that the best course of action would be to move the Option A neighborhoods forward to the standard-setting meetings without making any adjustments.

Policy Committee Surveys

Policy committee members were asked to complete two types of surveys during the course of the meeting: the neighborhood judgment readiness survey and the process evaluation survey. This section summarizes the outcomes of these surveys.

NEIGHBORHOOD JUDGMENT READINESS SURVEY

During each part of the committee judgment and feedback portion of the meeting, committee members were asked to fill out the neighborhood judgment readiness survey before providing their rank-ordering of the neighborhood options. The purpose of this survey was to confirm

that all committee members understood the empirical study results, that they understood what their judgment task was, and that they were ready to make their judgments. The readiness survey requested a "yes" or "no" response to the following statements:

- I understand my task for Part X (where X is 1, 2, or 3).
- I understand the data that were presented before Part X (where X is 1, 2, or 3).
- I am ready to begin Part X (where X is 1, 2, or 3).

Committee members recorded their unique panelist identification number on their survey so that the surveys could be collected and redistributed between each part of the meeting. Any committee member who was not ready to proceed with his/her judgments was directed to alert the facilitators. The facilitators would then answer questions or reexplain any information, concepts, or study results that were causing confusion.

A summary of the results from the neighborhood judgment readiness survey is provided in Table 6.13. All committee members indicated that they understood and were ready to proceed with their judgments in the three parts of the meeting.

Table 6.13: Summary of Policy Committee Neighborhood Judgment Readiness Survey

Readiness Statement	Part 1	Part 2	Part 3
Understood Task	100% Yes	100% Yes	100% Yes
Understood Data	100% Yes	100% Yes	100% Yes
Ready to Begin	100% Yes	100% Yes	100% Yes

PROCESS EVALUATION SURVEY

At the end of the policy committee meeting, committee members were asked to complete a process evaluation survey. The purpose of the process evaluation was to collect information about each committee member's experiences in recommending reasonable neighborhoods for the cut scores on the STAAR EOC assessments.

The survey was divided into five sections. The first section asked committee members to rate the successfulness of the various components of the policy committee meeting, such as the explanation of the purpose of the meeting and the background and requirements of the STAAR program, the discussion of the policy questions, and the presentation of the empirical studies and neighborhood options. The second section asked committee members to evaluate the adequacy of the amount of time spent on various elements of the meeting, such as the training, table discussions, and judgment tasks. In the third section, committee members were to provide their input on whether they thought that they were given adequate opportunities to express their professional opinions about policy questions and neighborhood options. The fourth section asked committee members whether they thought that they were provided adequate opportunities during the meeting to ask questions and interact with their fellow committee members. The fifth section was open-ended so that participants could provide additional comments about the process or their experience as a committee member. Panelists

were asked not to include any identifying information on the survey so that the responses would be anonymous.

A summary of the responses to the policy committee process evaluation survey is provided in Appendix 9. Most committee members thought that the various components of the meeting were "successful" or "very successful." They also thought that the time spent on training, table discussions, and judgment tasks was "adequate" to "very adequate." Virtually all committee members responded that they were given adequate opportunities to express their opinions about the policy questions and neighborhood options, to ask questions about the studies and neighborhoods, and to interact with other committee members.

STAAR 3–8 Empirical Studies

After the policy committee recommended the neighborhoods for the STAAR EOC assessments, the STAAR EOC standard-setting committees recommended cut scores for Levels II and III within the neighborhoods (see Chapter 7 for information on the standard-setting committees). The STAAR 3–8 neighborhood development guidelines were determined after the recommended performance standards for EOC were approved. The guidelines were developed based on empirical studies for STAAR 3–8, the STAAR EOC performance standards, and the EOC neighborhood guidelines. Table 6.14 lists the empirical studies that were use to prepare the neighborhood guidelines for STAAR 3–8 assessments.

Table 6.14: Validity and Linking 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

Unlike EOC, the data used in the empirical studies for STAAR 3–8 assessments were collected under motivated conditions during the spring 2012 administration, which was the first high-stakes administration. The representativeness of the data in terms of demographics and student proficiency was very similar to the student populations. For reading and mathematics, where assessments are offered at every grade level, the links across grades were based on large sample sizes, and the matching variables indicated strong relationships.

For science and writing, where assessments are not offered at every grade level, the span was three grade levels between the elementary and middle school assessments. Because of the three-year gap between assessments and the developmental differences between elementary and middle school students, the empirical studies for aligning the neighborhoods for writing and science were limited. This limitation required caution in the interpretation of the study results (see Chapter 3 for information on the empirical studies).

STAAR 3–8 Neighborhood Development

The recommendation from the policy committee to implement Neighborhood Option A, developed from the operational definitions and empirical study results for the EOC assessments, was also applied to the development of the STAAR 3–8 neighborhoods. In addition, in order to create an aligned system of performance standards, EOC performance standards were endorsed as anchors in establishing the STAAR 3–8 neighborhoods.

Texas Education Code (TEC) §39.0241 requires that performance standards be aligned from grade 3 through end-of-course assessments. Under an aligned set of standards, student performance at each level (i.e., Unsatisfactory, Satisfactory, or Advanced Academic Performance) within a content area should indicate whether or not the student is on track to be successful in the next grade or course. In order to align the performance standards in this way, TEA started with STAAR EOC assessments at the high school level and worked backward to grade 3 (see Figure 1.2 in Chapter 1). As such, once the performance standards for STAAR EOC assessments were determined, it was possible to establish neighborhoods for STAAR 3–8 assessments.

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

Figures 6.7 and 6.8 show the neighborhood development guidelines for STAAR 3–8. Figure 6.7 provides the details of the guidelines for each performance standard. Figure 6.8 illustrates the guidelines graphically.

Below Level II: Satisfactory
- ☒ Percent of test questions representing guessing
- ☒ TAKS Met Standard

Within the Neighborhood of Level II: Satisfactory
- ☒ Percent of students in Level II in their first EOC assessment in the same content area
- ☒ Reasonably likely (with at least a 60% probability) to succeed (reach Level II) in their first EOC assessment in the same content area
- ☒ NAEP Proficient or higher

Within the Neighborhood of Level III: Advanced
- ☒ Percent of students in Level III in their first EOC assessment in the same content area
- ☒ At least 50% probability of obtaining the EXPLORE College Readiness Benchmark
- ☒ At least 50% probability of obtaining the ReadiStep score aligned with the Texas postsecondary definition
- ☒ Highly likely (with at least a 75% probability) to succeed (reach Level II) in their first EOC assessment in the same content area
- ☒ NAEP Advanced or higher
- ☒ Within the region of precise measurement

Figure 6.7: STAAR 3–8 Neighborhood Development Guidelines

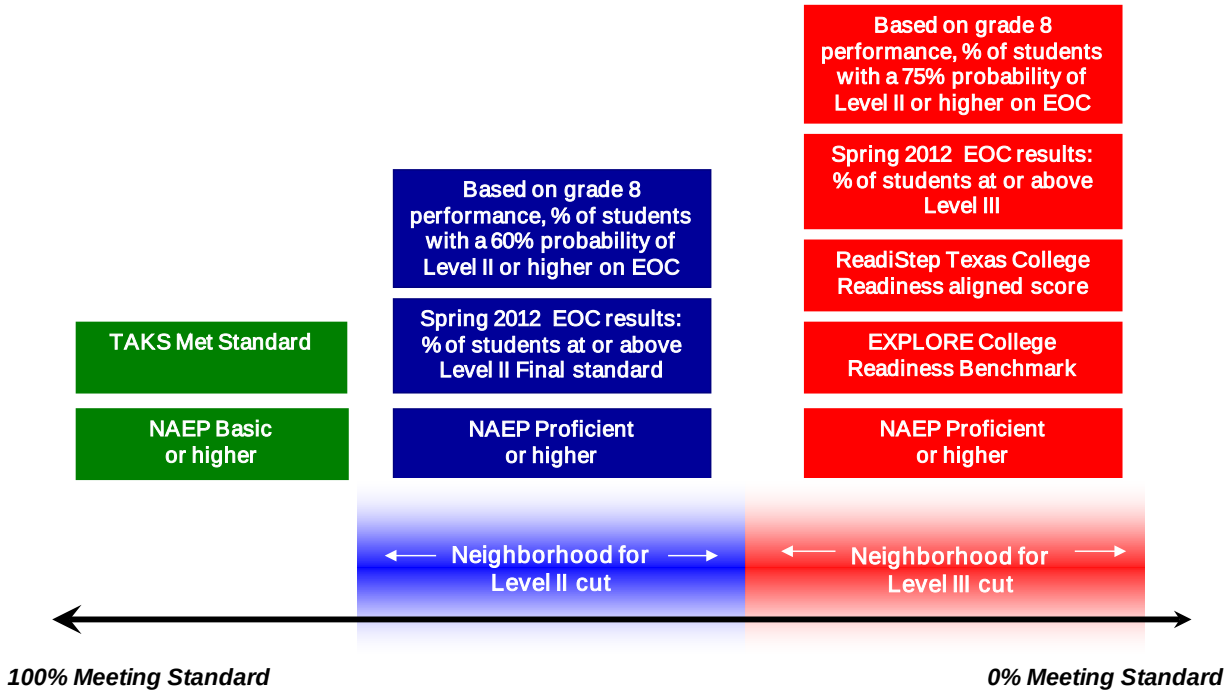


Figure 6.8: Graphical Illustration of STAAR 3–8 Neighborhood Development Guidelines

Empirical number lines, such as the one shown in Figure 6.9, were generated for all grade 8 assessments and grade 7 writing. The full set of empirical number lines used in determining the neighborhoods for grade 8 assessments and grade 7 writing is provided in Appendix 10. The impact data shown in the empirical number lines are based on student performance during the spring 2012 administration. The neighborhoods established for the STAAR grade 8 assessments

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.

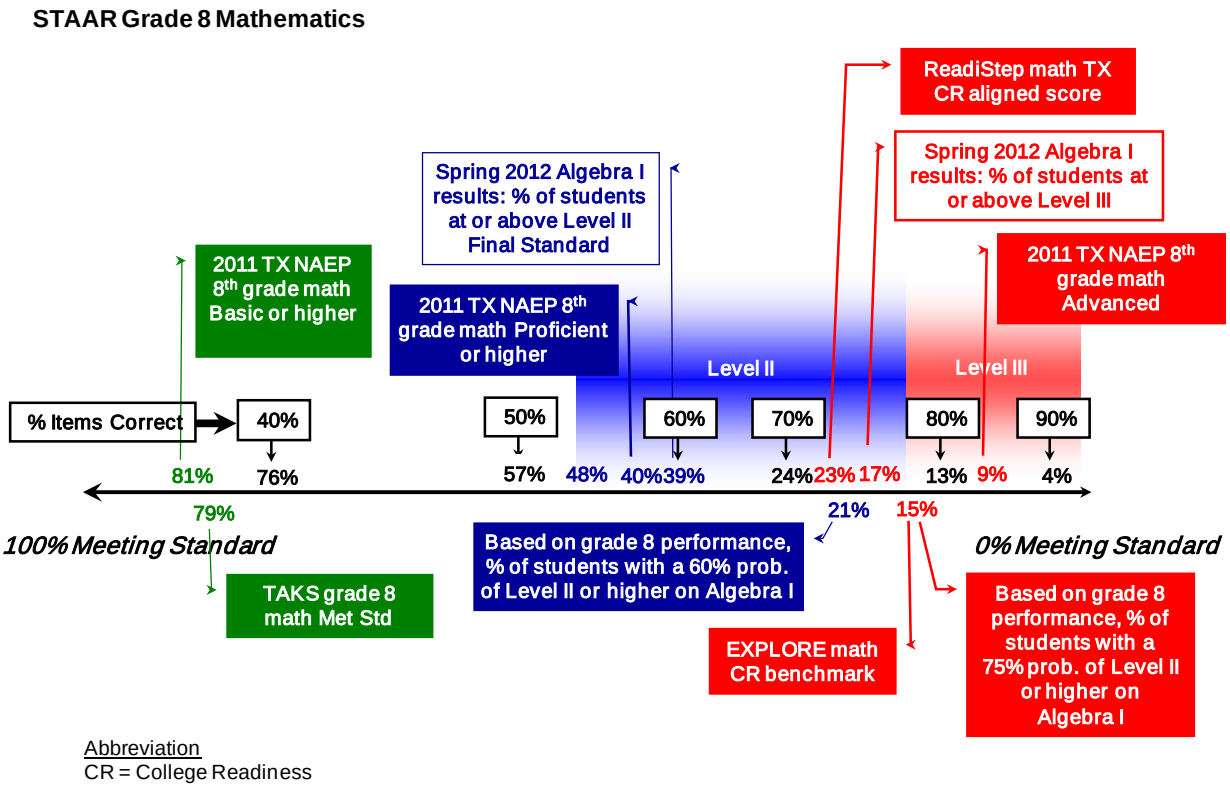


Figure 6.9: Empirical Number Line for STAAR Grade 8 Mathematics

Under TEC §39.036, TEA is required to develop a vertical scale in grades 3–8 for reading and mathematics. Because the vertical scales for reading and mathematics empirically link student performance on STAAR 3–8 assessments within the same subject area, 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 also informed the development of these neighborhoods.

Since the assessments within reading and mathematics were on the same vertical scale, number lines were not produced for each grade-level assessment. The neighborhoods were 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 6.10, were generated for grades 3–8 for reading and mathematics. The full set of vertical-scale neighborhoods is given in Appendix 12. The impact data shown in the vertical-scale graphic are based on student performance during the spring 2012 administration.

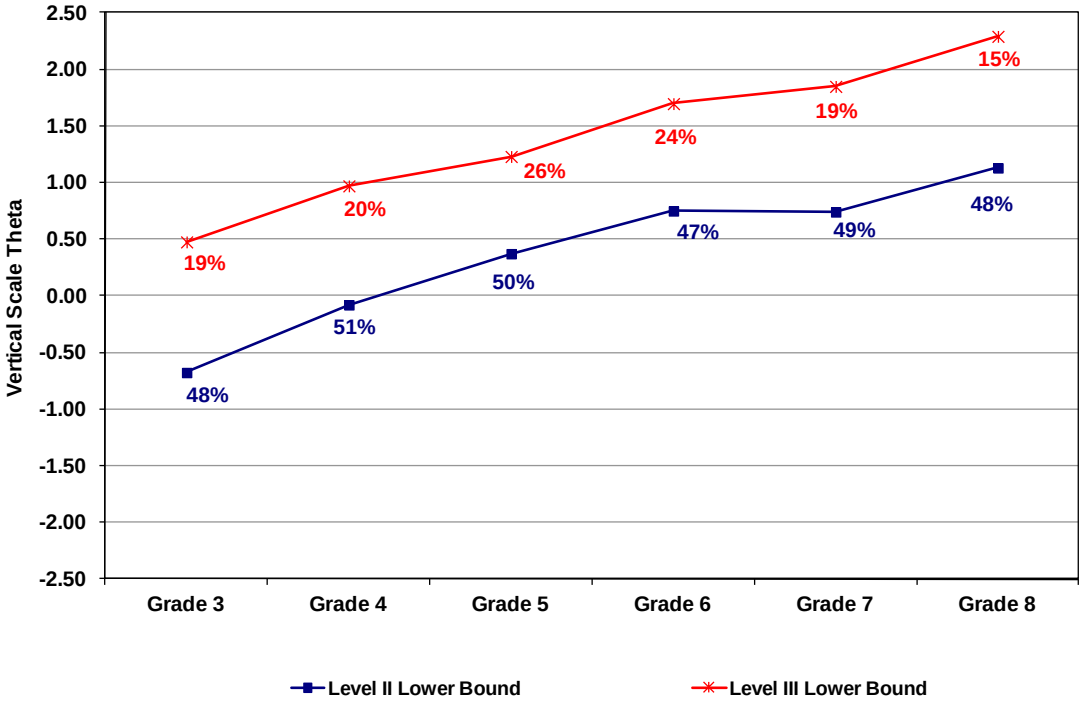


Figure 6.10: Vertical Scale Graphic for STAAR Grade 3–8 Mathematics Neighborhoods

For STAAR grade 4 writing and grade 5 science, a vertical scale was not available. The neighborhood number lines, such as the one shown in Figure 6.11, 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. These neighborhood number lines are provided in Appendix 11.

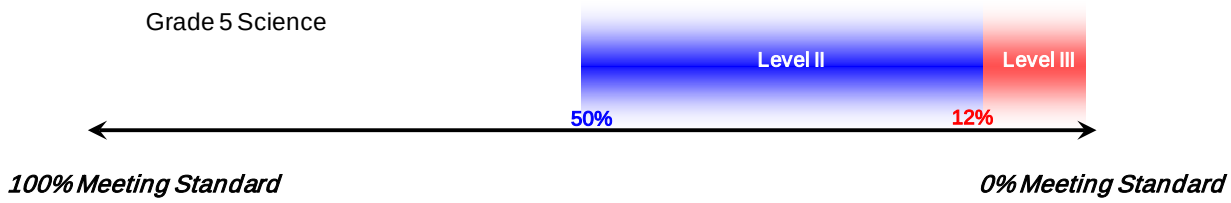


Figure 6.11: Example Neighborhood for STAAR Grade 5 Science

The neighborhoods established for the STAAR assessments provided reasonable ranges for the standard-setting committees to set performance standards. The next chapter provides more detail about the use of the neighborhoods in the standard-setting process.

Chapter 7: Standard-Setting Committees

This chapter provides details about Step 5 of the nine-step STAAR standard-setting process, which focuses on convening standard-setting meetings. The sections in this chapter include

- Purpose of Standard-Setting Committee Meetings
- Committee Composition and Attendees
- Description of the Standard-Setting Process
- Meeting Proceedings
- Recommended STAAR Cut Scores

Purpose of Standard-Setting Committee Meetings

All standard setting is based to a large degree on educator judgment. Panelists use their experience and knowledge to make expert recommendations. These judgments help establish the criteria for interpreting test scores using a specific standard-setting method. The purpose of holding STAAR standard-setting meetings was to gather expert recommendations for the performance standards on each STAAR assessment.

Each committee was asked to recommend cut scores for Level II: Satisfactory Academic Performance and Level III: Advanced Academic Performance using the following types of information:

- Content of the STAAR assessments
- Performance labels and policy definitions
- Performance Level Descriptors (PLDs) for each assessment
- Reasonable ranges (or neighborhoods) within which the cut scores should fall
- Selected results from empirical studies

Committee Composition and Attendees

When selecting standard-setting panelists, TEA placed an emphasis on content knowledge and classroom experience. However, the judgments and cut-score recommendations made by the committees were also guided by empirical studies, both through the neighborhoods and as feedback provided after each round of judgment.

Table 7.1 shows the groups from which panelists were recruited and the rationale for including each type of panelist.

Table 7.1: Recruitment Groups for Standard-Setting Committee Members

Recruitment Group	Rationale
Texas educators	Brought content knowledge and classroom experience from secondary and postsecondary institutions across Texas.
Special population representatives	Represented the perspective of English language learners (ELLs) and students served by special education. For grades 3–5 mathematics and grade 5 science, where one set of standards was set for both English and Spanish language tests, the committees included educators who were familiar with the differences in instruction in classrooms where both English and Spanish languages are used and with the specific needs of students in these situations.

To help the standard-setting committees gain an understanding of the steps that took place before these meetings, TEA included some panelists who also served on the specific PLD committees and could share their experiences. In addition, several policy-committee members attended the EOC standard-setting meetings as observers so that they could see how the neighborhood recommendations were used.

The tables in Appendix 13 summarize the characteristics and experience of the panelists on each standard-setting committee. These tables provide demographic information about the committee members as well as information about the members' current positions in education, the number of years they have been in their positions, their experience working with the various types of student populations, and the types of districts they represent.

Description of the Standard-Setting Process

The evidence-based standard-setting approach (Beimers, Way, McClarty, & Miles, 2012; O'Malley, Keng, & Miles, 2012) was used to set performance standards on the STAAR assessments. This approach incorporated features of several different standard-setting methods. Elements of the benchmark method (Phillips, 2011) were included by using the bookmark (or item-mapping) method with external data (Ferrara, Lewis, Mercado, D'Brot, Barth, & Egan, 2011). Ordered item booklets (OIBs) used by the standard-setting committees were created based on the policy committee's neighborhood recommendations. (See below for more information about OIB development.) 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. Between the judgment rounds, the panelists were provided information—including empirical study results and impact data—that they used to refine their judgments. By suggesting that panelists place a bookmark within a neighborhood, the variation among the panelists' judgments were limited, resulting in performance standards that were reasonable based on the empirical studies.

The STAAR standard-setting process incorporated a review of various data sources (as was done in Haertel's briefing book approach [2002, 2012]), including empirical studies such as SAT and ACT external validity studies, STAAR-to-TAKS comparison studies, STAAR vertical scaling studies for reading and mathematics in grades 3–8, and a contrasting-groups study (Livingston & Zieky, 1982) using student performance in entry-level college courses to distinguish between students who were successful in their college courses and those who were not. The results of these studies were used during the policy-committee meeting to guide the development of the neighborhoods. During the standard-setting committee meetings, the information from the various studies was presented, along with the neighborhoods, to guide the panelists' judgments regarding appropriate performance-level cut scores.

PROCESS FOR DEVELOPING ORDERED ITEM BOOKLETS (OIBS)

Panelists received instructions in item mapping. The item-mapping procedure required panelists to review a set of test questions, or items, and decide which of them were likely to be answered correctly by students just barely within a given performance level. Each OIB's test items were ordered from easiest to most difficult (see Figure 7.1). As the items became progressively more difficult, panelists decided, item by item, whether a student just barely within a performance level would be likely to respond correctly.

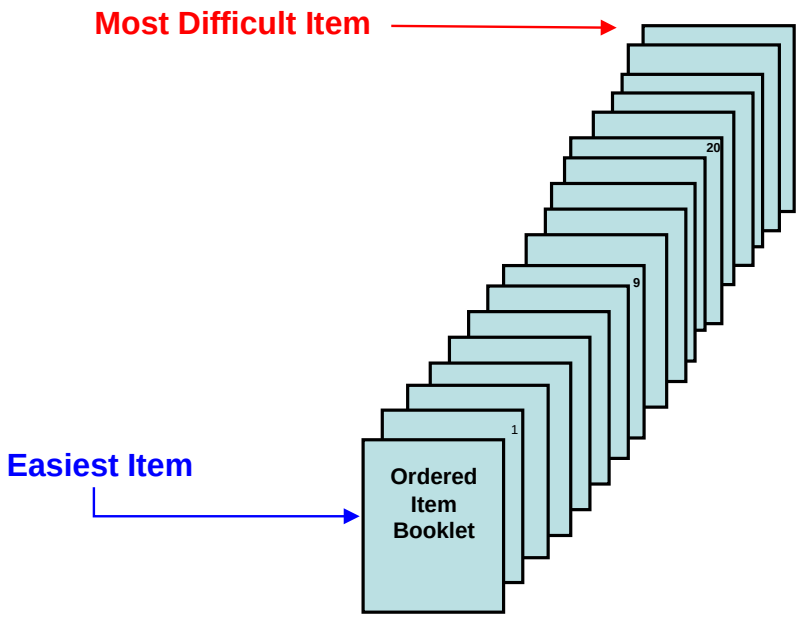


Figure 7.1: Arrangement of Items in an Ordered Item Booklet

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. After being administered to students, test items were calibrated using the Rasch model (Rasch, 1980) to obtain Rasch item difficulty values. These values were used to order the items from easiest to most difficult in the OIBs. 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.

Once neighborhoods were determined for the cut scores, each OIB was evaluated to make sure that the full scale range of the assessment was represented by the items in the OIB. Areas of the OIB that did not have item representation along the 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.

The item-mapping procedure with a response probability (RP) value of 0.67 was used to create the OIBs and facilitate panelist judgments for all STAAR mathematics, science, and social studies meetings and for the grades 3–8 reading meetings. That is, items were mapped to the difficulty scale at the point at which students had roughly a 2/3 probability of answering the item correctly. However, a slightly different procedure was used to create the OIBs and facilitate panelist judgments for the STAAR grades 4 and 7 writing and the English I, II, and III meetings. The STAAR grades 4 and 7 writing and the English I, II, and III writing assessments include written compositions that require students to generate a response. The English I, II, and III reading assessments include short-answer reading items that similarly require students to generate a response. To create the OIBs using the item-mapping procedure, multiple short-answer and written composition items would have been required in order to fill gaps in the neighborhood ranges. Adding multiple short-answer and written composition items to the OIBs may have been confusing to panelists and would have exposed a large number of items that could be easily memorized. To avoid these issues, the yes/no procedure (Impara & Plake, 1997) was used to create the OIBs and guide panelist judgments for the STAAR grades 4 and 7 writing and the English I, II, and III assessments. Phillips (2002) discusses the use of the yes/no procedure for TAKS standard setting. When the RP value of 0.67 was used, gaps in the OIB meant that there could be places within the STAAR scale range where panelists could not place a cut-score recommendation because there were no items in the OIB representing that point on the STAAR scale range. The yes/no procedure allowed for the OIBs to be created with a wide enough range of items within the neighborhoods without the inclusion of additional short-answer and written composition items. By using this approach to the OIBs, panelists were able to place cut-score recommendations across the full range of the STAAR scale.

Meeting Proceedings

For the STAAR EOC assessments, standard-setting meetings were conducted during a two-week period. Because of the content overlap in the English reading, English writing, and mathematics EOC assessments, one committee for each content area (reading, writing, and mathematics) recommended standards for all assessments in that content area. The following committees met for three days on February 22–24, 2012:

1. English reading committee—English I reading, English II reading, and English III reading
2. English writing committee—English I writing, English II writing, and English III writing
3. Mathematics committee—Algebra I, geometry, and Algebra II

Six separate committees were convened to recommend standards for each of the STAAR EOC science and social studies assessments. The following committees met for 2.5 days on February 29–March 2, 2012:

1. Biology
2. Chemistry
3. Physics
4. World geography
5. World history
6. U.S. history

For the STAAR 3–8 assessments, sixteen standard-setting meetings were conducted during a two-week period in October 2012 and an additional day in November 2012 to recommend standards for 21 assessments. Each committee met for approximately two days. The organization of the standard-setting meetings allowed the recommended standards from the upper grade-level committees to be used as feedback for the lower grade-level committees. Table 7.2 lists the committees based on the subjects and grades and the meeting dates.

Table 7.2: STAAR 3–8 Standard-Setting Committee Meeting Organization

Dates	Committees by Subjects and Grades
10/2/2012 – 10/3/2012	Mathematics grade 8
	Reading grade 8
	Writing grade 7
	Science grade 8
	Social studies grade 8
10/4/2012 – 10/5/2012	Mathematics grades 6 and 7
	Reading grades 6 and 7
10/9/2012 – 10/10/2012	Mathematics grade 5
	English reading grade 5
	Spanish reading grade 5
	Science grade 5
10/11/2012 – 10/12/2012	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

STAAR grades 4 and 7 writing committees met for an additional day in November to consider updated information regarding the OIBs, impact data, and neighborhoods.

Committees recommended a total of 72 cut scores (30 for STAAR EOC and 42 for STAAR 3–8) — two cut scores (Level II: Satisfactory Academic Performance and Level III: Advanced Academic Performance) for each of the STAAR assessments. Table 7.3 shows the agenda for the standard-setting committee meetings.

Table 7.3: Standard-Setting Committee Meeting Agenda

General Session	• Welcome and Introductions • Background Information • Overview of Standard Setting
Breakout Sessions	• Specific Performance Level Descriptors* • Borderline Students* • Standard-Setting Training • Round 1: Judgment and Feedback* • Round 2: Judgment and Feedback* • Round 3: Judgment and Feedback* • Cross-Course Articulation (EOC) or Group Discussion (3–8) • Evaluation and Closing Remarks

* These tasks were repeated for each assessment for which the committee was recommending standards.

Figure 7.2 gives an overview of the sequence of events that was followed for the STAAR EOC standard-setting meetings. For English and mathematics, the events differed slightly from those for science and social studies as a result of the difference in meeting format.

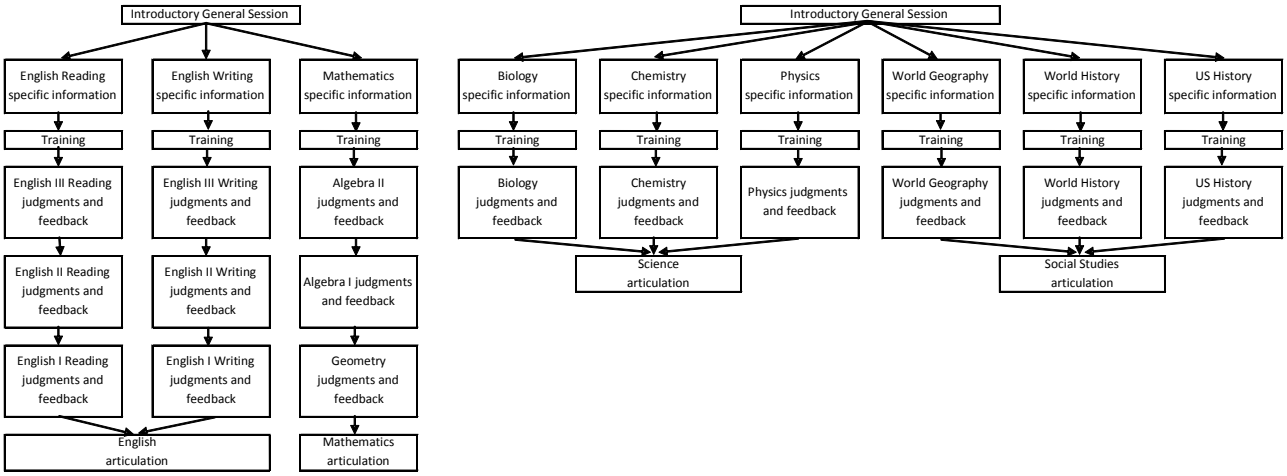


Figure 7.2: Overview of the STAAR EOC Standard-Setting Committee Meetings

The STAAR 3–8 standard-setting meetings generally followed the format of the STAAR EOC meetings, except that there was no cross-course articulation. Instead, each committee had the opportunity to review their recommendations in conjunction with the recommendations of committees that met previously (i.e., higher grade levels). Figure 7.3 gives an overview of the sequence of events for three examples: the grade 8 reading committee meeting, the grades 6 and 7 reading committee meeting, and the grade 5 reading committee meeting. For the grades 3–5 reading and grade 4 writing assessments, the English and Spanish committees met together for the specific PLDs discussion, the borderline students discussion, and the standard-setting training. The committees separated before judgments and feedback occurred.

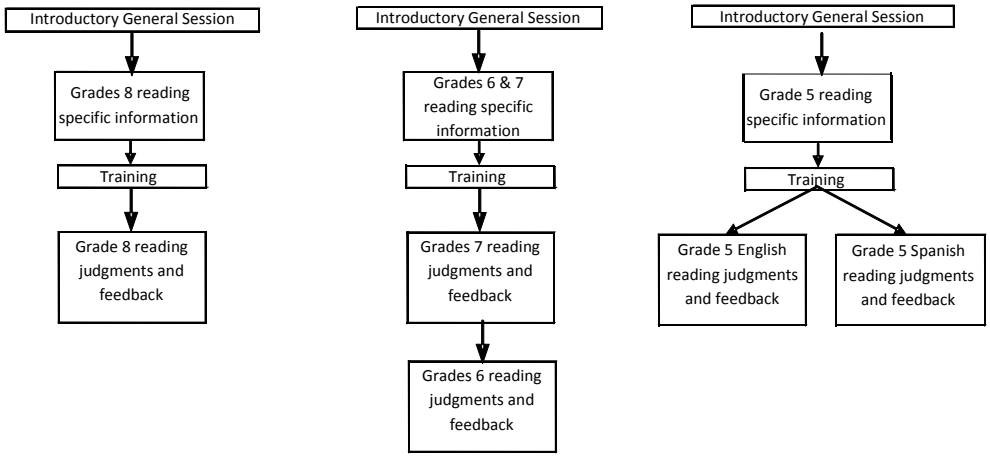


Figure 7.3: Overview of the STAAR 3–8 Standard-Setting Committee Meetings

A description of each topic in the agenda is provided next.

GENERAL SESSION

The purpose of the general session was to welcome the standard-setting committees; give background information about legislative requirements, the STAAR program, and standard setting; and describe the standard-setting committee's responsibilities. A single general session took place at the beginning of each standard setting.

WELCOME AND INTRODUCTIONS

TEA welcomed the panelists. Key facilitators of the standard-setting process were introduced, and general housekeeping tasks were covered, including the non-disclosure agreement, security protocols, and reimbursement forms. Committee members were introduced once panelists had moved to their breakout sessions.

BACKGROUND INFORMATION

The STAAR standard-setting panelists were provided with a review of the STAAR program, including information on legislative requirements, graduation plans and requirements, test eligibility, time limits, and accommodations. Policy considerations and plans for phasing in the standards were also discussed with all STAAR panelists. The nine-step process being used to establish performance standards for STAAR was shared, including an overview of how the

Performance Descriptor Advisory Committee helped develop the performance labels and policy definitions and how reasonable ranges (or neighborhoods) were developed within which the standard-setting committee should work in order to recommend standards.

OVERVIEW OF STANDARD SETTING

To help panelists understand what standard setting is and the reason they were asked to be part of a standard-setting committee, facilitators discussed the purpose of setting standards and provided panelists with information about the specific standard-setting approach being used during the meetings.

BREAKOUT SESSIONS

After the general session, panelists moved into their content-specific breakout sessions for the remainder of the meeting. Within each committee, panelists were divided into three to four table groups. Each table group consisted of different types of committee members so that there was a blend of expertise at each table. Table leaders were identified to facilitate the discussions and assist in meeting logistics (for example, by collecting judgment forms) at each table.

Information specific to the assessment(s) for which each committee was setting standards was then presented. Each committee member had the opportunity to take a sample STAAR assessment. The goal was for each committee member to see what a test form looks like and get a sense of the types of items and content included on STAAR as well as the depth of knowledge required on it. After taking a sample assessment, panelists checked their responses and discussed the test-taking experience. In addition, panelists discussed the assessment itself in terms of content, difficulty, and the construct being measured.

PERFORMANCE LEVEL DESCRIPTORS (PLDS)

To help inform discussions, facilitators directed panelists to review the performance labels, policy definitions, and PLDs for each assessment. The PLDs are a framework for a common understanding of the knowledge, skills, and abilities possessed by a student at each performance level (Level III: Advanced Academic Performance, Level II: Satisfactory Academic Performance, and Level I: Unsatisfactory Academic Performance). The PLDs gave the panelists guidance about what students should know and be able to do within each performance level for a specific STAAR assessment. When reviewing the PLDs, panelists were asked to think about what most differentiates Level III: Advanced Academic Performance from Level II: Satisfactory Academic Performance and Level II: Satisfactory Academic Performance from Level I: Unsatisfactory Academic Performance.

BORDERLINE STUDENTS

After reviewing the PLDs, panelists were asked to think about the group of students who would just barely reach Level II: Satisfactory Academic Performance and the group of students who would just barely reach Level III: Advanced Academic Performance. These are the “borderline” students—defined as those students who have the minimum amount of knowledge necessary to be in Level II or Level III. Panelists were asked to work in their table groups to come up with

descriptors that characterize what a borderline student for Level II and Level III should know and be able to do. Whereas the PLDs described the student in the middle of a performance level, the borderline descriptors focused on students with just enough knowledge to get them into a performance level. Table groups shared their borderline descriptors, and the committee as a whole discussed each group's contribution in order to develop a master set of borderline descriptors for Level II and Level III for each assessment. These descriptors were used by panelists while making their judgments throughout the remainder of the meeting.

STANDARD-SETTING TRAINING

The committee members received training on the bookmark, or item-mapping, procedure (Lewis, Mitzel, Green & Patz, 1999) for mathematics, science, social studies, and the grades 3–8 reading assessments, and training on the yes/no procedure (Impara & Plake, 1997) for the STAAR grades 4 and 7 writing, English I, English II, and English III assessments. Panelists used these procedures to recommend the cut scores for each assessment. For both processes, the OIB was a primary tool. Panelists practiced evaluating items and making cut-score recommendations using an abbreviated "practice" OIB in order to try out the item-mapping and yes/no procedures. Before making judgments, panelists were asked to read each item, identify the knowledge and skills needed for a correct response, and review the PLDs.

As they made their judgments, panelists were asked to think about the borderline student and the descriptors they had previously developed. For the item-mapping procedure, panelists were asked to look through the "practice" OIB and identify the last item that a borderline student would have a 2/3 probability of answering correctly. A marker was placed on the last item that a borderline student had a 2/3 chance of answering correctly. For the yes/no procedure, panelists were asked to look through the "practice" OIB and identify the last item that they expected a borderline student would answer correctly. A marker was placed on the last item that a borderline student would answer correctly. After the practice session, the group discussed any questions or difficulties related to the mechanics of the procedures.

Panelists were given information about the purpose of the neighborhoods and the use of validity studies to determine the neighborhoods. When panelists received their actual OIB, they were instructed to mark the lower boundary of the Level II: Satisfactory Academic Performance neighborhood with a blue flag. Panelists were asked to place their first-round Level II recommendations after the blue flag. Panelists were also asked to look through all the items appearing before the blue flag in the OIB and were given the option of placing their marker before the blue flag if they had a strong rationale for doing so. The lower boundary of Level III was not shown to panelists until after they had placed their Level II marker.

Once the panelists placed their Level II cut-score recommendations, they were asked to place a red flag in the OIB to indicate the lower bound of the Level III: Advanced Academic Performance neighborhood. Panelists were asked to place their first-round Level III recommendations after the red flag unless they had a strong rationale for doing otherwise.

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. Before each judgment round, panelists were asked to complete a readiness form indicating their understanding of the tasks required of them and their readiness to participate in the next round.

Even though the item-mapping and yes/no procedures were used by the committee to recommend its cut scores, the STAAR standard-setting process as a whole incorporated aspects of a number of established standard-setting methods. These methods included the use of empirical validity evidence and feedback from policy stakeholders, as recommended in the briefing book method (Haertel 1999; 2002); application of the contrasting-groups method (Zieky & Livingston, 1977) as part of the empirical studies informing standard setting; and the concept of rangefinding and pinpointing to determine the neighborhoods and cut scores, a process commonly found in portfolio-based standard-setting methods.

After receiving training on the item-mapping or yes/no procedure, the committee members participated in three rounds of judgments for each assessment. For the meetings where one committee set standards for multiple assessments, the committee started with the cut-score recommendations in the most advanced grade or course and worked backward. This approach was in line with the goal of making STAAR a comprehensive system, with performance standards that are aligned to and link back from postsecondary readiness to high school to middle school to elementary school.

Within each round, panelists were asked to consider the items in the OIB, starting with the easiest item. Each panelist made 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.

ROUND 1: JUDGMENT AND FEEDBACK

During the first round of judgments, committee members made their cut-score recommendations primarily based on the content of the OIB and the neighborhood ranges identified within the OIB. After the Round 1 judgments, the following types of feedback were presented:

- 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 STAAR EOC assessments, 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 external measures of postsecondary readiness or to the next course in a course-taking sequence.
- 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.

An example of committee-level Round 1 feedback can be found in Appendix 14.

For the English III reading, English III writing, and Algebra II assessments, the data included the probability of passing an entry-level general education college course in the same content area (based on results from the "college students take STAAR" study) and projected SAT and/or ACT scores in the related content area (based on results from the study that compared STAAR with SAT and/or ACT).

For the other EOC English and mathematics assessments, the data included the likelihood of attaining Level II or III on the next test in the content area (based on results from the STAAR linking studies). For the STAAR EOC science assessments, the empirical data included the projected ACT science scores. Since ACT and SAT do not include a specific social studies assessment, these projections were not available to the social studies committees. Reference information, such as the average ACT and SAT scores for students enrolled in various Texas colleges and the ACT college-readiness benchmark scores, was provided to panelists as a point of reference for the feedback related to external tests.

For the grade 8 assessments and the grade 7 writing assessment, the data included on-track information that indicates the students' probability of success (reaching Level II) on the next test typically taken within the content area. For the grade 8 reading, mathematics, and science assessments and the grade 7 writing assessment, the data included the likelihood of reaching the EXPLORE benchmark in the related content area. Additionally, the grade 8 mathematics test also included the likelihood of reaching the ReadiStep benchmark in mathematics. The likelihood of reaching the ReadiStep benchmark was not shared with the grade 8 reading and grade 7 writing committees since the results of these linking studies were outside the reasonable ranges. Therefore, the likelihood of reaching these benchmarks would always be 100% regardless of where the committee made their judgments. Since EXPLORE and ReadiStep do not include a specific social studies assessment, these likelihoods were not available to the grade 8 social studies committee.

Information that related scores on the STAAR assessments with outside 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 right 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. In addition, 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 associated with their bookmark placement.

For the grades 3–7 reading and mathematics assessments, the committee was shown a graph of the vertical scale that contained their Level II and Level III cut-score recommendations in conjunction with the higher grade-level cut-score recommendations from the prior committees in the same subject area. An example of vertical scale feedback can be found in Appendix 14. 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.

ROUND 2: JUDGMENT AND FEEDBACK

For the second round of judgments, committee members made their cut-score recommendations based on the first-round feedback, discussion with their table groups, and content of the items in the OIB. After completing their Round 2 judgments, panelists were provided with the following second-round feedback:

- The panelist's individual Round 2 cut-score recommendations (bookmarked pages) for Level II and Level III
- Table-level Round 2 cut-score recommendations—the minimum, maximum, mean, and median bookmarked pages for Level II and Level III
- Committee-level Round 2 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 2 cut-score recommendation
- Impact data for the test based on the committee's Round 2 cut-score recommendations
- For STAAR EOC assessments, feedback data included NAEP impact data
- For STAAR EOC assessments, grade 8 assessments, and the grade 7 writing assessment, additional feedback included data showing projections from the committee-level Round 2 cut-score recommendations to external measures of postsecondary readiness or to the next course in a course-taking sequence.
- For STAAR grades 3–7 reading and mathematics assessments, additional feedback included vertical scale data showing the committee-level Round 2 cut-score recommendations and the higher grade-level Level II and Level III cut-score recommendations from prior committees.

The impact data for each STAAR EOC assessment were based on performance during the spring 2011 administration. Panelists were cautioned that the spring 2011 administration was a low-stakes administration for students and that the corresponding impact data might reflect a lack of motivation by students taking the assessments. The impact data for each STAAR 3–8 assessment were based on performance during the spring 2012 administration, which was the

first high-stakes administration of STAAR. An example of committee-level Round 2 feedback can be found in Appendix 14.

ROUND 3: JUDGMENT AND FEEDBACK

During the third round of judgments, committee members made their final individual cut-score recommendations based on all the feedback they received in the first two rounds. The feedback each panelist received after Round 3 included the following:

- The panelist's individual Round 3 cut-score recommendations (bookmarked pages) for Level II and Level III
- Table-level Round 3 cut-score recommendations—the minimum, maximum, mean, and median bookmarked pages for Level II and Level III
- Committee-level Round 3 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 3 cut-score recommendation
- Impact data for the test based on the committee's Round 3 cut-score recommendations
- For STAAR EOC assessments, feedback data included NAEP impact data.
- For STAAR EOC assessments, grade 8 assessments, and the grade 7 writing assessment, additional feedback included data showing projections from the committee-level Round 3 cut-score recommendations to external measures of postsecondary readiness or to the next course in a course-taking sequence.
- For STAAR grades 3–7 reading and mathematics assessments, additional feedback included vertical scale data showing the committee-level Round 3 cut-score recommendations and the higher grade-level Level II and Level III cut-score recommendations from prior committees.

The feedback from Round 3 was used as the primary input to the cross-course articulation process and the group discussion described below for the STAAR assessments.

STAAR EOC CROSS-COURSE ARTICULATION

The final activity in which the STAAR EOC standard-setting panelists participated was the cross-course articulation. The purpose of the cross-course articulation was to look at the cut-score recommendations (presented as page numbers in the OIB) that were made across all STAAR EOC assessments in a content area and evaluate the reasonableness of these cuts. Panelists were shown the impact data resulting from their Round 3 cut-score recommendations across all courses within a content area. Recommendations for cut-score adjustments could be made by the committee as a group after reviewing the Round 3 feedback and group discussion. Any recommended changes made during the cross-course articulation had to be supported by a review of the OIB and the PLDs for that assessment. The committee was also asked to review and refine the PLDs as necessary so that there was solid alignment between the final committee cut-score recommendations and the PLDs.

STAAR 3–8 Group Discussion

The final activity for the STAAR 3–8 standard-setting panelists was a group discussion among all the panelists. The purpose of the group discussion was to give the standard-setting committee the opportunity to address the reasonableness of their cut-score recommendations after receiving Round 3 feedback. Recommendations for cut-score adjustments could be made by the committee as a group after reviewing the Round 3 feedback and group discussion. Any recommended changes made during the group discussion had to be supported by a review of the OIB and the PLDs for that assessment. The committee was also asked to review and refine the PLDs as necessary so that there was solid alignment between the final committee cut-score recommendations and the PLDs.

Process-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. The seven-part survey asked committee members to provide feedback on

1. The level of success of the various components of the meeting
2. The usefulness of the activities conducted during the meeting
3. The adequacy of the various components of the meeting
4. How confident committee members were that the PLDs accurately reflected student performance at each performance level
5. How confident committee members were about the final cut-score recommendations
6. Whether committee members thought that they had been given adequate opportunities to express their professional opinions, ask questions, and interact with others
7. Whether committee members thought that their judgments and opinions had been respected by their fellow panelists and by the facilitators

Panelists were asked not to include any identifying information on the survey so that their responses would be anonymous.

Although there was some variation across committees, most committee members 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. A summary of the responses to the standard-setting committee process evaluation is provided in Appendix 15.

Recommended STAAR Cut Scores

Recommendations Resulting from Judgment Rounds

The cut-score recommendations resulting from each round of judgment are presented (in terms of OIB page number) in Appendix 16. Although there were some Level II cut-score recommendations placed within the Level III neighborhoods (past the red flag), after the third round of judgments, 69 of the 72 median cut-score recommendations were within the Level II and Level III neighborhoods provided. One Level II cut-score recommendation was within the Level III neighborhood, and two Level III cut-score recommendations were above the Level III neighborhood.

Descriptive statistics (including the minimum, maximum, standard deviation, mean, and median cut-score recommendations) for each round of judgment can be found in Appendix 17. Graphical representations of data regarding panelist agreement across rounds can be found in Appendix 18. In general, variation across panelist judgments decreased across rounds.

Recommendations Resulting from STAAR EOC Cross-Course Articulation

The Round 3 impact data shown during the cross-course articulation or group discussion and impact data resulting from changes made during the articulation or group discussion can be found in Appendix 19. The actual page-number recommendations resulting from the articulation can be found in Appendix 16. Table 7.4 shows the changes that were recommended during the articulation of each content area and the rationale that the committee used to support the change.

Table 7.4: Cross-Course Articulation Recommendations

Content Area	STAAR Assessment	Articulation Result
Mathematics Articulation	Algebra I	The Level III cut score was raised to align better with the Algebra II Level III cut score.
	Geometry	The Level II cut score was lowered because some committee members thought the Round 3 cut score was too high, although not all members agreed about this recommendation.
	Algebra II	No change was made.
English Articulation	English I Reading	No change was made.
	English II Reading	No change was made.
	English III Reading	The Level II cut score was raised to align better with the cut scores on English I and II reading.
	English I Writing	No change was made.
	English II Writing	No change was made.
	English III Writing	The Level II cut score was raised to align better with the cut scores on English I and II writing.

Table 7.4 cont.: Cross-Course Articulation Recommendations

Content Area	STAAR Assessment	Articulation Result
Science Articulation	Biology	No change was made.
	Chemistry	No change was made.
	Physics	The Level II cut score was lowered, and the Level III cut was raised; however, the committee did not reach consensus on either of these recommendations.
Social Studies Articulation	World Geography	No change was made.
	World History	No change was made.
	U.S. History	No change was made.

RECOMMENDATIONS RESULTING FROM STAAR 3–8 GROUP DISCUSSION

The actual page-number recommendations resulting from the group discussions can be found in Appendix 16. For the STAAR 3–8 assessments, only one change was recommended during the group discussion. The Level III cut score for the grade 8 reading assessment was raised because the committee members thought the Round 3 cut score was too low based on the items in the OIB.

The recommendations from the standard-setting committees were then reviewed by TEA as part of the reasonableness review (see Chapter 8). Committee discussions that occurred throughout the standard-setting meetings were used in determining the final cut scores provided to the Commissioner of Education and, in the case of STAAR Algebra II and English III, the Commissioner of Higher Education.

Chapter 8: Reasonableness Review

This chapter provides details about Step 6 in the nine-step STAAR standard-setting process, which focuses on reviewing performance standards for reasonableness. The sections in this chapter include the following:

- Purpose of Reasonableness Review
- Rationale for Adjustments Made During Reasonableness Review
- Reasonableness Review Results

Purpose of Reasonableness Review

After educator committees recommended Level II and Level III performance standards for the STAAR assessments, TEA conducted reasonableness reviews of the cut-score recommendations across content areas and made adjustments as appropriate. The reasonableness review process following standard setting is intended to confirm that performance standards contribute to a well-articulated and coherent assessment system. During the STAAR EOC cross-course articulations, the standard-setting committees were able to evaluate the results of their judgments within a content area. The reasonableness review was conducted not only within a content area but also across content areas in order to evaluate how well the standards across all STAAR assessments aligned with one another. For the STAAR 3–8 assessments, TEA 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.

During the reasonableness review, TEA evaluated the results from all the standard-setting committees (see Appendix 16). The following pieces of information were considered:

- The content discussions that occurred during the standard-setting meetings
- The alignment of content expectations to the performance level descriptors and policy definitions, both within and across content areas
- Round 3 judgments of the standard-setting committees, including
    - The content of items in the OIB that were close to the cut scores recommended during Round 3
    - The impact data associated with the Round 3 judgments
    - The empirical study results associated with Round 3 judgments
- Changes recommended during the cross-course articulation or group discussion, including
    - The content of items in the OIB that were close to the cut scores recommended during the articulation or group discussion
    - The impact data associated with the articulation or group discussion recommendations
    - The empirical study results associated with the articulation or group discussion recommendations

The review of this information was used to determine the final cut-score recommendations for Level II: Satisfactory Academic Performance and Level III: Advanced Academic Performance.

Rationale for Adjustments Made During Reasonableness Review

During the reasonableness review, the results of the cross-course articulation and group discussions were evaluated first. Cut-score recommendations were reviewed across content areas to look for inconsistencies in content alignment to performance levels, impact data, and validity study results. The following types of questions were evaluated as part of the review.

- Did the skills required to correctly answer the items in the OIB that were close to the cut-score recommendations truly reflect the knowledge and skills of students at the Level II and Level III performance levels?
- Did the impact data, or percentage of students estimated to achieve Level II and Level III performance, seem reasonable for each content area?
- Did the recommended cut scores make sense given the empirical study results and the performance level descriptors?

In addition, a critical feature of the STAAR assessment system is the alignment of performance standards across sequential assessments within a content area (for example, Algebra I and Algebra II).

Panelists' discussions during the standard-setting meetings were taken into consideration, especially if there was a lack of consensus among committee members. If there was significant disagreement among committee members and the recommended cut scores seemed misaligned when looked at as a system, the OIB was reviewed to confirm that the item content (of items close to the marked page in the OIB) matched the performance level descriptors for the performance level. Table 8.1 shows the changes that were recommended during the STAAR EOC reasonableness review of each content area and the rationale used to support the change. Table 8.2 shows the changes that were recommended during the STAAR 3–8 reasonableness review of each content area and the rationale used to support the change.

Table 8.1: STAAR EOC Reasonableness Review Recommendations

Content Area	STAAR Assessment	Recommendation after Reasonableness Review
Mathematics	Algebra I	Both the Level II and Level III cut scores for Algebra II were set quite high. During articulation, the standard-setting committee shifted the Level III Algebra I cut score higher to better align with the Level III Algebra II cut score. However, the Level II Algebra I cut score was not moved during articulation, even though some committee members thought it should be moved. By not moving the Level II Algebra I cut score, there was a misalignment between the performance expectations on these two mathematics assessments. Therefore, it was recommended that the Level II Algebra I cut score be raised.

Table 8.1 cont.: STAAR EOC Reasonableness Review Recommendations

Content Area	STAAR Assessment	Recommendation after Reasonableness Review
Mathematics	Geometry	During articulation, the Level II cut score was lowered because some committee members thought the Round 3 cut score was too high, but not all members agreed on this recommendation. When viewed with the other mathematics assessments and assessments in other content areas, the Level II cut score seemed too low. In fact, the Level II geometry cut score from Round 3 was more in alignment with the other mathematics assessments and was the same cut score that had been recommended from both Round 1 and Round 2. For these reasons, it was recommended that the Level II geometry cut-score recommendation be the Round 3 recommendation from the standard-setting committee.
	Algebra II	No change; use articulation results
English	English I Reading	No change; use articulation results
	English II Reading	No change; use articulation results
	English III Reading	No change; use articulation results
	English I Writing	No change; use articulation results
	English II Writing	No change; use articulation results
	English III Writing	No change; use articulation results
Science	Biology	No change; use articulation results
	Chemistry	No change; use articulation results
	Physics	During the articulation process, the physics committee was unable to reach consensus. The articulation results represented a compromise, but committee members were told that all viewpoints would be used when evaluating the final cut-score recommendations during reasonableness review. In addition, the impact data did not align well with the impact data from the cut scores recommended for chemistry. Based on the reasonableness review, the Level II physics cut- score recommendation was raised from the articulation result to the recommendation from Round 3. This allowed for better alignment with chemistry. In addition, the Level III physics cut-score recommendation was raised to better align with the Level III recommendation for chemistry and to better align with the pattern of results seen in other content areas.

Table 8.1 cont.: STAAR EOC Reasonableness Review Recommendations

Content Area	STAAR Assessment	Recommendation after Reasonableness Review
Social Studies	World Geography	When reviewed, the Level II cut scores for world geography did not seem to have a strong alignment between the content expectations for items close to the cut score and the Level II performance level descriptors. Though the committee did not make a change during articulation, there was discussion that higher cut scores could be supported in anticipation of adjustments in instruction that would match the assessed content. For these reasons the Level II world geography cut-score recommendation was raised to be more in line with the world history and U.S. history cut-score recommendations.
	World History	No change; use articulation results
	U.S. History	No change; use articulation results

Table 8.2: STAAR 3–8 Reasonableness Review Recommendations

Content Area	STAAR Assessment	Recommendation after Reasonableness Review
Mathematics	Grade 3	When reviewed, the Level III cut-score recommendation did not align well across grades along the vertical scale, or in terms of impact data, with most other mathematics Level III cut-score recommendations. Based on the reasonableness review, the Level III grade 3 mathematics cut-score recommendation was lowered to align with the cut-score recommendations for grades 4–8. There was no change to the Level II recommended standard.
	Grade 4	No change; use group discussion results
	Grade 5	When reviewed, the Level III cut-score recommendation did not align well across grades along the vertical scale, or in terms of impact data, with most other mathematics Level III cut-score recommendations. Based on the reasonableness review, the Level III grade 5 mathematics cut-score recommendation was lowered to the committee's Round 1 recommendation to bring the grade 5 cut-score recommendation into alignment. There was no change to the Level II recommended standard.
	Grade 6	No change; use group discussion results
	Grade 7	No change; use group discussion results
	Grade 8	No change; use group discussion results

Table 8.2 cont.: STAAR 3–8 Reasonableness Review Recommendations

Content Area	STAAR Assessment	Recommendation after Reasonableness Review
Reading	Grade 3 English	The Level II and Level III cut scores were set very close together by the committee. In order to have an appropriate distance between the Level II and Level III cut scores, given the standard error of measurement of the test forms, the Level II cut-score recommendation was lowered to the committee's Round 1 recommendation. There was no change to the Level III recommended standard.
	Grade 4 English	No change; use group discussion results
	Grade 5 English	No change; use group discussion results
	Grade 6 English	No change; use group discussion results
	Grade 7 English	No change; use group discussion results
	Grade 8 English	No change; use group discussion results
	Grade 3 Spanish	No change; use group discussion results
	Grade 4 Spanish	No change; use group discussion results
	Grade 5 Spanish	No change; use group discussion results
Science	Grade 5	No change; use group discussion results
	Grade 8	No change; use group discussion results
Social Studies	Grade 8	No change; use group discussion results
Writing	Grade 4 English	When reviewed, the impact data for Level II did not align well with the impact data for cut-scores recommended for other writing assessments. The Level II cut-score recommendation was raised to be more in alignment with the other writing assessments. The alignment of cut scores for Level II across writing assessments will inform the development of coherent writing programs from elementary school to middle school to high school that support students' acquisition of the writing skills needed to write effectively on-grade level. There was no change to the Level III recommended standard.
	Grade 4 Spanish	No change; use group discussion results
	Grade 7	No change; use group discussion results

Reasonableness Review Results

Of the 72 cut scores recommended by the standard-setting committees, adjustments were made to nine cut scores:

- Mathematics
  - Grade 3 — Level III: Advanced Academic Performance
  - Grade 5 — Level III: Advanced Academic Performance
  - Algebra I — Level II: Satisfactory Academic Performance
  - Geometry — Level II: Satisfactory Academic Performance

- Reading
    - Grade 3 English — Level II: Satisfactory Academic Performance
- Science
    - Physics — Level II: Satisfactory Academic Performance
    - Physics — Level III: Advanced Academic Performance`
- Social Studies
    - World geography — Level II: Satisfactory Academic Performance
- Writing
    - Grade 4 English — Level II: Satisfactory Academic Performance

A summary of the standard-setting committees' Round 3 cut-score recommendations, the results of the cross-course articulations and group discussions, and all adjustments implemented during the reasonableness review can be found in Appendix 16. The impact data associated with each set of recommendations can be found in Appendix 19.

Chapter 9: Approval of Performance Standards

This chapter provides details about Step 7 of the nine-step STAAR standard-setting process, which is focused on approving performance standards. The sections in this chapter include:

- Determination of Phase-in Cut Scores
- Establishing Phase-in and Minimum Scores for STAAR EOC Assessments
- Establishing Phase-in Scores for STAAR 3–8 Assessments
- Final Approval of Recommended, Phase-in, and Minimum Scores

Determination of Phase-in Cut Scores

Implementing phased-in performance standards is not new to Texas. When standards for the TAKS program were set in 2002, three performance categories were established: Commended Performance, Met Standard, and Did Not Meet Standard. For TAKS, a two-year phase-in period was established for the Met Standard performance standard. During the first year of TAKS implementation (2003 for grades 3–10 and 2004 for grade 11 exit level), a cut score was set two standard errors of measurement (SEMs) below the panel-recommended cut score for Met Standard. During the second year of TAKS implementation, that standard increased to one SEM below the panel-recommended Met Standard cut score. The panel-recommended standard was then implemented in the third year.

A number of options similar to the TAKS phase-in plan were considered for the STAAR assessments. However, the change from a high school grade-based assessment system to a course-based assessment system introduced complexities to the phase-in process. For example, the EOC cumulative score requirement within a content area would be difficult to determine if students had a different phase-in standard for each assessment taken within a content area. This could make it difficult to know whether a student should retest.

Because of these complexities, STAAR EOC phase-in standards will not be applied according to grade-year student cohorts. Rather, each Texas student will be held to phase-in standards based on the first time a course is taken within a content area, and that standard will be held constant within academic content areas throughout his or her high school career. Specifically, if a student takes his first STAAR EOC assessment in 2013 or before, he will be held to the first set of phase-in performance standards for every assessment in that content area. Likewise, if a student takes her first STAAR EOC assessment in 2014 or 2015, she will be held to the second set of phase-in performance standards.

The STAAR 3–8 phase-in standards will be applied based on administration year since the same complexities do not exist at grades 3–8.

A variety of empirical studies were conducted to help inform the recommended Level II and Level III performance standards for STAAR assessments (see Chapter 3). Likewise, this research was used to inform the phase-in standards and minimum scores. The STAAR empirical studies

established statistical linkages between performance on STAAR assessments and complementary measures such as high school and college course grades and admissions test scores. These linkages were in turn used to inform reasonable ranges on STAAR scales within which phase-in cut scores could be set. The use of empirical studies to inform phase-in standards for STAAR EOC and STAAR 3–8 are discussed in the following sections.

Establishing Phase-in and Minimum Scores for STAAR EOC Assessments

Phase-in cut scores were determined empirically for each STAAR assessment after the reasonableness review of the performance standards was completed. The following three sections discuss the phase-in cut scores for STAAR EOC Level II performance standards, Level III performance standards, and minimum scores.

Level II: Satisfactory Academic Performance

The final cut-score recommendations that came from the standard-setting committees and the reasonableness review were at the upper end of the neighborhood boundaries established by the policy committee. For this reason, many of the study points that were conceptually within the Level II neighborhood going into standard setting fell below the final recommended standards. Figure 9.1 displays the study points that map above, below, and close to the STAAR EOC phase-in Level II performance standards.

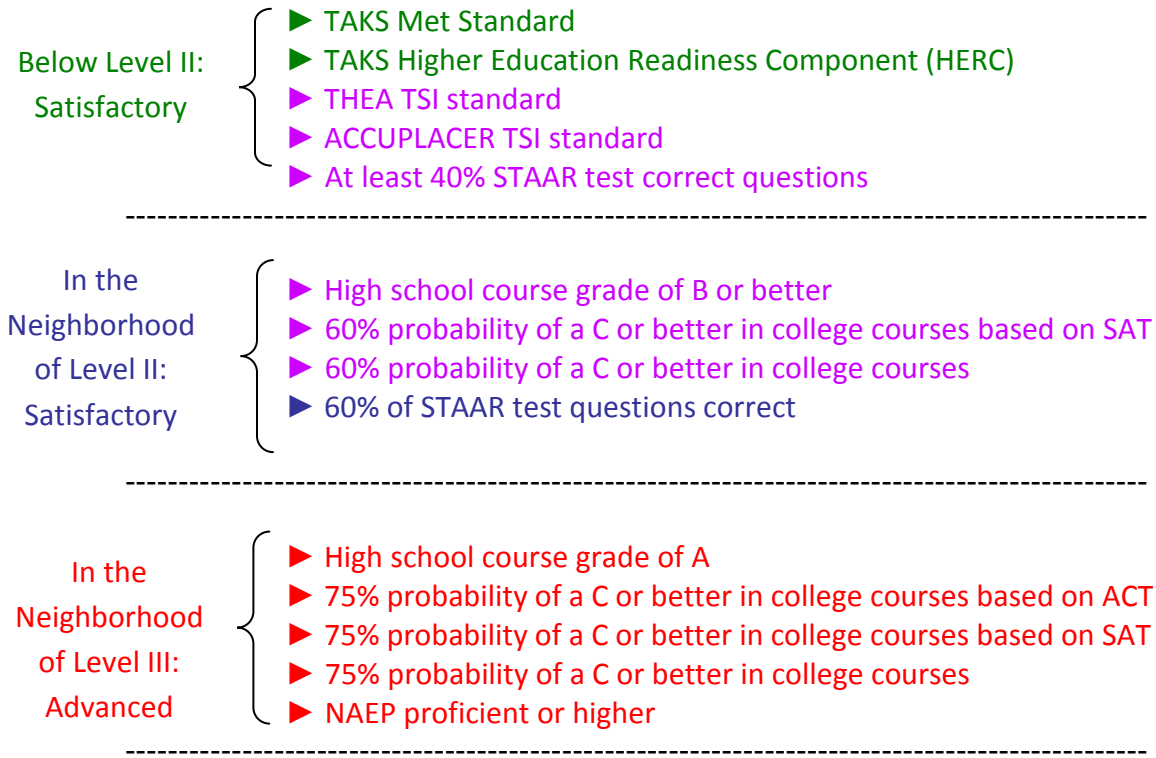


Figure 9.1: Empirical Study Considerations for Phase-in Level II Performance Standards

In Figure 9.1, study mappings in purple font highlight the region where phase-in standards could reasonably be established. These study mappings could then be used to guide the placement of the phase-in standards. This figure is adapted from information used to guide recommended Level II and Level III performance standards. It is important to note the alignment of these phase-in adjustments with empirical study evidence. Each of the adjusted phase-in standards achieves the following:

1. The phase-in Level II standards for STAAR EOC assessments are higher than the TAKS Met Standard.
2. The phase-in Level II standards for STAAR EOC assessments are higher than the TAKS HERC cut scores.
3. The phase-in Level II standards for Algebra II, English III reading, and English III writing are all higher than the THEA score required for enrolling in entry-level, credit-bearing college courses.
4. Across STAAR EOC assessments, adjustments based on standard deviations align with the validity study mappings in Figure 9.1, which are intended to inform reasonable phase-in cut scores.

Because of the significant increase in rigor of the STAAR program, the distance between the phase-in and final standards for STAAR Level II is generally twice as large as the distance was between phase-in and final standards under the TAKS Met Standard. Therefore, the STAAR Level II phase-in occurs over a four-year rather than a two-year period.

A four-year, two-step phase-in for Level II was implemented in order to provide school districts with an appropriate amount of time to adjust instruction, provide new professional development, increase teacher effectiveness, and close knowledge gaps. During this time period, both the phase-in and recommended standards will be reported. In fall 2014, recommended standards will be reviewed (see Chapter 11) and possibly changed based on additional validity studies and student scores from motivated, high-stakes assessment conditions. The Level II phase-in approach is illustrated in Table 9.1, where typical STAAR test-taking sequences are presented for multiple cohorts of Texas students.

In Table 9.1, STAAR assessments in bold indicate that phase-in performance standards will be applied; STAAR assessments not in bold indicate that recommended performance standards will be applied. The vertical dashed green and blue lines mark the beginning of the first and second phase-in periods, respectively, for Level II. The vertical dashed orange line signals the implementation of the final recommended performance standards for Level II.

Table 9.1: STAAR Phase-in of Performance Standards Across Multiple Cohorts of Students

Level II* Phase-in for All STAAR Assessments

Cohort	2011–2012	2012–2013	2013–2014	2014–2015	2015–2016	2016–2017	2017–2018
1	Grade 9 **Algebra I**	Grade 10 **Geometry**	Grade 11 **Algebra II**				
2	Grade 8 **Mathematics**	Grade 9 **Algebra I**	Grade 10 **Geometry**	Grade 11 **Algebra II**			
3	Grade 7 **Mathematics**	Grade 8 **Mathematics**	Grade 9 **Algebra I**	Grade 10 **Geometry**	Grade 11 **Algebra II**		
4	Grade 6 **Mathematics**	Grade 7 **Mathematics**	Grade 8 **Mathematics**	Grade 9 **Algebra I**	Grade 10 **Geometry**	Grade 11 **Algebra II**	
5	Grade 5 **Mathematics**	Grade 6 **Mathematics**	Grade 7 **Mathematics**	Grade 8 **Mathematics**	Grade 9 Algebra I	Grade 10 Geometry	Grade 11 Algebra II
6	Grade 4 **Mathematics**	Grade 5 **Mathematics**	Grade 6 **Mathematics**	Grade 7 **Mathematics**	Grade 8 Mathematics	Grade 9 Algebra I	Grade 10 Geometry
7	Grade 3 **Mathematics**	Grade 4 **Mathematics**	Grade 5 **Mathematics**	Grade 6 **Mathematics**	Grade 7 Mathematics	Grade 8 Mathematics	Grade 9 Algebra I

Level III Phase-in for STAAR Algebra II, English III Reading, and English III Writing**

Cohort	2011–2012	2012–2013	2013–2014	2014–2015	2015–2016	2016–2017	2017–2018
1	Grade 9 Algebra I	Grade 10 Geometry	Grade 11 **Algebra II**				
2	Grade 8 Mathematics	Grade 9 Algebra I	Grade 10 Geometry	Grade 11 **Algebra II**			
3	Grade 7 Mathematics	Grade 8 Mathematics	Grade 9 Algebra I	Grade 10 Geometry	Grade 11 Algebra II		
4	Grade 6 Mathematics	Grade 7 Mathematics	Grade 8 Mathematics	Grade 9 Algebra I	Grade 10 Geometry	Grade 11 Algebra II	
5	Grade 5 Mathematics	Grade 6 Mathematics	Grade 7 Mathematics	Grade 8 Mathematics	Grade 9 Algebra I	Grade 10 Geometry	Grade 11 Algebra II

*The Level II phase-in example used above will be applied to all STAAR assessments.

**The Level III phase-in example used above will be applied only to Algebra II, English III reading, and English III writing. There is no phase-in of Level III for the other STAAR assessments.

A variety of measures (for example, standard error of measurement, conditional standard error of measurement, standard deviation, distance between STAAR and TAKS standards) were considered in order to establish phase-in Level II standards relative to recommended Level II standards. As noted previously, TAKS employed a standard error of measurement as the unit by which Met Standard cuts were phased in. The options for the unit of measure to use for the STAAR phase-in were shared with the Texas Technical Advisory Committee (TTAC) on August 25–26, 2011, March 22–23, 2012, and September 27–28, 2012. Feedback from the TTAC indicated that the decision to use a specific unit of measure for phase-in should be based on empirical evidence, have a strong policy rationale, and include an evaluation of the impact data resulting from the phase-in standard. In the case of STAAR assessments, the standard deviation in scale score units provides phase-in standards that (1) are supported by empirical measures,

(2) meet the current instructional needs of Texas students and districts and provide time to adjust instruction, and (3) support scale-score cuts that are consistent across tests within a content area. More specifically, the following adjustments (in scale-score units) were implemented to set phase-in Level II cut scores:

- For 2012 and 2013
  - –0.5 standard deviations for STAAR EOC English reading and writing assessments
  - –1.0 standard deviation for all STAAR EOC mathematics, science, and social studies assessments
- For 2014 and 2015
  - –0.2 standard deviations for STAAR EOC English reading and writing assessments
  - –0.5 standard deviation for all STAAR EOC mathematics, science, and social studies assessments

Multiple comparisons using STAAR and TAKS scores suggested that the distance between phase-in standards and final standards for the STAAR program is generally greater than the distance between the phase-in standards and final standards for the TAKS program. However, external study data indicated that a slightly smaller phase-in is appropriate for the English reading and writing STAAR EOC assessments compared to the phase-in required for the STAAR EOC mathematics, science, and social studies.

For the most part, satisfying the multiple priorities of the phase-in plan could be accomplished via a one-standard-deviation adjustment. For English reading and writing STAAR EOC assessments, that adjustment was modified to 0.5 standard deviations. This change was implemented to maintain alignment of English phase-in standards with empirical evidence.

LEVEL III: ADVANCED ACADEMIC PERFORMANCE (ENGLISH III AND ALGEBRA II)

An evidence-based approach to setting phase-in cut scores was also used for Level III. Some of the studies noted in Figure 9.1 are available only for those EOC assessments designed specifically to provide postsecondary-readiness indicators—STAAR Algebra II, English III reading, and English III writing. Specifically, Level III phase-in cut scores were set at the point on each STAAR EOC scale where students have at least a 75% likelihood of earning a grade of C or better in an entry-level college course. These mappings are based on the "college students taking STAAR" study and the SAT and ACT external validity studies (see Chapter 3 for more information about the empirical studies). Furthermore, for English III reading, English III writing, and Algebra II, the phase-in cut scores for Level III are higher than both the recommended Level II performance standards and the corresponding TAKS Commended Performance cut scores.

A two-year, one-step phase-in was implemented for Level III for English III reading, English III writing, and Algebra II. The Level III performance standards for the other STAAR assessments were not phased in. Students need to achieve Level III performance on English III reading,

English III writing, and Algebra II in order to be eligible to graduate under the Distinguished Achievement Program. Because the distance between the phase-in standard and final standard for Level III is smaller than the distance for Level II, a two-year, one-step phase-in is appropriate.

Like Level II, the STAAR EOC Level III phase-in will also begin on a student-by-student basis: if a student takes his or her first EOC assessment within a content area in 2013 or before, he or she will be held to the phase-in Level III performance standard for English III and Algebra II. However, if a student takes his first EOC assessment within a content area in 2014 or later, he will be held to the final recommended Level III standard. The Level III phase-in approach is illustrated in Table 9.1. The vertical dashed orange line signals the implementation of the final recommended performance standards for Level III.

MINIMUM SCORES

In addition to the recommended and phase-in performance standards, minimum scores were set empirically to fall below the phase-in and recommended Level II cut scores. Scores on STAAR EOC assessments that fall below the minimum score cannot be counted toward a student's cumulative score, which is required for graduation. Cumulative score requirements in the phase-in period are calculated as the product of the phase-in Level II standard and the number of assessments in that content area. Students must reach cumulative score requirements in each content area as one part of earning a high school diploma.

Minimum scores for the STAAR EOC program were established using the conditional standard error of measurement (CSEM). This statistic, commonly used when Rasch models are applied, varies across the performance continuum and across separate assessments (Embretson & Reise, 2000). The CSEM is defined as the reciprocal of the test information function (TI) at the level of performance corresponding to the Level II cut score (θ_{Level_II}):

$$CSEM(\theta_{Level\_II}) = \frac{1}{\sqrt{T1(\theta_{Level\_II})}}$$

The recommended minimum score falls one CSEM below the recommended Level II performance standard. Likewise, each phase-in minimum score falls one CSEM below the relevant phase-in Level II performance standard. Therefore, whether recommended or phase-in standards are being implemented, the minimum score is always set one CSEM below the Level II cut score. The rationale underlying the use of the CSEM to set minimum scores is focused on measurement error. While a student's score may fall in Level I, his or her score may be close enough to the Level II performance standard that the difference could be attributed to measurement error around the cut score. The CSEM captures this concept. In addition, the CSEM — unlike the classical standard error of measurement — accounts for varying degrees of measurement precision across the performance continuum.

Establishing Phase-in Scores for STAAR 3–8 Assessments

Phase-in cut scores were determined empirically for each STAAR assessment after the reasonableness review of the performance standards was completed. The following section discusses the phase-in cut scores for STAAR 3–8.

LEVEL II: SATISFACTORY ACADEMIC PERFORMANCE

For STAAR 3–8, the empirical studies that informed the phase-in Level II performance standards included the STAAR-to-TAKS comparisons, the linking studies from STAAR 3–8 to EOC, and the vertical scale study. The linking studies informed the alignment of the phase-in performance standards between STAAR 3–8 and EOC. The vertical scale informed the alignment of the phase-in standards across tests within a content area for STAAR 3–8 reading and mathematics to support increasing performance standards from grades 3–8. The phase-in cut scores were set lower than recommended Level II standards but still higher than the TAKS Met Standard and guessing.

The STAAR 3–8 phase-in began with the 2012 administration, and similar to STAAR EOC, a four-year, two-step phase-in for Level II was implemented. Phase-in 1 cut scores for Level II will be in effect for the 2012 and 2013 administrations. Phase-in 2 cut scores for Level II will be in effect for the 2014 and 2015 administrations. The final recommended Level II standards will be in place for STAAR 3–8 assessments in 2016 and beyond. No phase-in will be applied to Level III.

The adjustments (in scale score units) to the Level II cut scores for STAAR 3–8 assessments are based on the same rationale as the STAAR EOC adjustments. For 2012 and 2013, the Level II phase-in 1 adjustments were –1.0 standard deviation for all STAAR 3–8 assessments. For 2014 and 2015, the Level II phase-in 2 adjustments were –0.5 standard deviation for all STAAR 3–8 assessments.

Final Approval of the Recommended, Phase-in, and Minimum Scores

In April 2012, the STAAR EOC performance standards were approved by the Commissioner of Education. Standards approval included not only Level II and Level III phase-in and recommended cut-scores but also minimum scores for each STAAR EOC assessment. Table 9.2 presents the approved standards and minimum scores in scale score units for each STAAR EOC assessment.

In December 2012, the STAAR 3–8 performance standards were approved by the Commissioner of Education. Standards approval included Level II and Level III recommended cut-scores and Level II phase-in cut-scores. Table 9.3 presents the approved standards in scale score units for each STAAR 3–8 assessment.

Table 9.2: STAAR EOC Performance Standards and Minimum Scores

Assessment	Phase-in 1 Minimum	Phase-in 1 Level II	Phase-in 2 Minimum	Phase-in 2 Level II	Final Recommended Minimum	Final Recommended Level II	Phase-in Level III	Final Recommended Level III
English I Reading	1813	1875	1887	1950	1936	2000	N/A	2304
English II Reading	1806	1875	1880	1950	1929	2000	N/A	2328
English III Reading	1808	1875	1882	1950	1932	2000	2135	2356
English I Writing	1798	1875	1872	1950	1921	2000	N/A	2476
English II Writing	1807	1875	1880	1950	1928	2000	N/A	2408
English III Writing	1808	1875	1881	1950	1929	2000	2155	2300
Algebra I	3371	3500	3626	3750	3872	4000	N/A	4333
Algebra II	3350	3500	3604	3750	3852	4000	4080	4411
Geometry	3362	3500	3619	3750	3868	4000	N/A	4397
Biology	3367	3500	3621	3750	3868	4000	N/A	4576
Chemistry	3348	3500	3600	3750	3846	4000	N/A	4607
Physics	3346	3500	3600	3750	3848	4000	N/A	4499
World Geography	3383	3500	3632	3750	3874	4000	N/A	4404
World History	3326	3500	3576	3750	3822	4000	N/A	4634
U.S. History	3372	3500	3624	3750	3869	4000	N/A	4440

Table 9.3: STAAR 3–8 Performance Standards

Assessment	Phase-in 1 Level II	Phase-in 2 Level II	Final Recommended Level II	Final Recommended Level III
Grade 3 English Mathematics	1392	1460	1529	1615
Grade 4 English Mathematics	1471	1535	1599	1677
Grade 5 English Mathematics	1489	1558	1627	1710
Grade 6 Mathematics	1509	1584	1658	1762
Grade 7 Mathematics	1551	1615	1678	1798
Grade 8 Mathematics	1583	1641	1700	1863
Grade 3 English Reading	1331	1400	1468	1555
Grade 4 English Reading	1422	1486	1550	1633
Grade 5 English Reading	1458	1520	1582	1667
Grade 6 Reading	1504	1567	1629	1718
Grade 7 Reading	1556	1615	1674	1753
Grade 8 Reading	1575	1637	1700	1783
Grade 4 English Writing	3500	3750	4000	4612
Grade 7 Writing	3500	3750	4000	4602
Grade 5 English Science	3500	3750	4000	4402
Grade 8 Science	3500	3750	4000	4406
Grade 8 Social Studies	3500	3750	4000	4268

Assessment	Phase-in 1 Level II	Phase-in 2 Level II	Final Recommended Level II	Final Recommended Level III
Grade 3 Spanish Mathematics	1392	1460	1529	1615
Grade 4 Spanish Mathematics	1471	1535	1599	1677
Grade 5 Spanish Mathematics	1489	1558	1627	1710
Grade 3 Spanish Reading	1304	1374	1444	1532
Grade 4 Spanish Reading	1398	1469	1539	1636
Grade 5 Spanish Reading	1447	1515	1582	1701
Grade 4 Spanish Writing	3500	3750	4000	4543
Grade 5 Spanish Science	3500	3750	4000	4402

Chapter 10: Implementation of Performance Standards

This chapter provides details about Step 8 of the nine-step STAAR standard-setting process, which is focused on implementing performance standards. Specifically, it provides a description of the STAAR scaling methodology, that is, the means by which performance on a STAAR assessment is converted to a scale score. The sections in this chapter include the following:

- STAAR EOC Scale Score System
- STAAR 3–8 Scale Score System
- Scaling Constants
- Rounding Rules

STAAR EOC Scale Score System

Implementing STAAR performance standards required converting Rasch-based performance estimates (θ) to scale scores, which are reported to districts, campuses, parents, and students. Generally, the approach to developing the STAAR EOC scale score system was driven by two priorities.

1. Within each content area, establish consistent scale score values corresponding to Level II: Satisfactory Academic Performance standards.
2. Within each content area, establish consistent standard deviations for each assessment's scale.

The first priority is intended to create a scale-score system as simple and consistent as possible for students, parents, campuses, and school districts. Because scale scores corresponding to the Level II cut are consistent within content areas, calculation of the cumulative score target within a content area is straightforward.

- For mathematics, science, and social studies assessments, the cumulative score target is equivalent to the Level II scale score cut multiplied by three, which is the number of assessments students are required to take in each content area under the RHSP or DAP.
- For English assessments, the cumulative score target is equivalent to the Level II scale score cut multiplied by six, which is the number of assessments students are required to take under the RHSP or DAP.

The second priority serves two purposes. First, as noted in Chapter 9, phase-in performance standards were established in standard deviation units at a fixed distance from recommended Level II cut scores. When each assessment in a content area has an equivalent standard deviation in scale score units, all phase-in scale score cuts within that content area will be consistent. Second, when scale scores are summed to satisfy a cumulative score requirement,

consistent standard deviations across assessments ensure that each assessment will carry the same weight in the cumulative score calculation.

STAAR 3–8 Scale-Score System

The STAAR 3–8 assessments have two scale-score systems, vertical scales and horizontal scales, based on the content area. A vertical scale refers to a conversion of a raw score onto a scale that is common to all assessments that measure a similar content area (e.g., mathematics) across different grades. With a vertical scale, a student's scale score in one grade can be directly compared to that student's scale score in another grade, making it possible to determine how much the student has progressed in that content area. The main advantage of a vertical scale is the ability to interpret year-to-year progress as demonstrated by scale-score changes. The reading and mathematics assessments are reported on vertical scales, since these content areas are assessed in adjacent grades from grade 3 to grade 8.

A horizontal scale converts a raw score onto a scale that allows for comparisons across test forms from year to year for a specific assessment. As with vertical scales, horizontal scales maintain the passing standard that students are required to meet in order to reach Level II or Level III performance categories. Unlike vertical scales, horizontal scale scores cannot be compared to scale scores for other grades in a content area.

VERTICAL SCALE SCORES

Under TEC § 39.036, TEA is required to develop a vertical scale for assessing student performance in grades 3–8 for reading and mathematics. A vertical scale was developed for the following grades and subjects:

- STAAR English grades 3–8 reading
- STAAR English grades 3–8 mathematics
- STAAR Spanish grades 3–5 reading

The vertical scale established for the English version of grades 3–5 mathematics was applied to the Spanish versions of grades 3–5 mathematics, since the Spanish versions of the mathematics tests are transadapted from the English test forms. A vertical scale is not required for science, social studies, or writing at the elementary or middle school level.

Scaling Constants

HORIZONTAL SCALING CONSTANTS

STAAR scale scores represent linear transformations of Rasch-based performance estimates (θ). Specifically, this transformation is accomplished by first multiplying any given θ by a slope (A) and subsequently adding an intercept (B). This simple operation is given by the equation below:

$$SC_\theta = A \times \theta + B \quad (1)$$

The slope (*A*) and intercept (*B*) in Equation (1) are referred to as scaling constants, and they are derived via a method described by Kolen and Brennan (2004). Two parameters describing the desired scale score system are established a priori: a scale score equivalent — a specific score point value on the STAAR scale — and the standard deviation of the scale. As noted above, controlling these characteristics (the scale score at the Level II cut and the standard deviation of each scale) were two priorities in standards implementation. The *A* scaling constant is calculated as follows:

$$A = \frac{\sigma_{sc}}{\sigma_{\theta}} \tag{2}$$

In Equation (2), σ_{sc} represents the desired standard deviation of the scale, while σ_{θ} represents the standard deviation of Rasch-based θ values among a sample group. To construct the STAAR EOC scales, the sample group for a given STAAR EOC assessment consisted of all students who took that assessment in 2011. For the STAAR 3–8 scales, the sample group for a given STAAR 3–8 assessment consisted of all students who took that assessment in spring 2012.

The *B* scaling constant is calculated as follows:

$$B = SC_{Level\_II} - \frac{\sigma_{sc}}{\sigma_{\theta}} \times \theta_{Level\_II} \tag{3}$$

In Equation (3), SC_{Level_II} represents the desired scale score at the Level II cut, and θ_{Level_II} represents the approved Level II performance standard (in Rasch units). As in Equation (2), σ_{sc} represents the desired standard deviation of the scale, while σ_{θ} represents the standard deviation of Rasch-based θ values in the sample group. Using Equation (1) and substituting Equation (2) for *A* and Equation (3) for *B*, the full STAAR scaling equation is shown below.

$$SC_{\theta} = \frac{\sigma_{sc}}{\sigma_{\theta}} \times \theta + \left[SC_{Level\_II} - \frac{\sigma_{sc}}{\sigma_{\theta}} \times \theta_{Level\_II} \right] \tag{4}$$

Because each STAAR assessment's scale is derived using its own sample group, σ_{θ} varies across assessments. Likewise, each assessment has a unique Level II performance standard in Rasch units, so θ_{Level_II} varies across assessments. SC_{Level_II} and σ_{sc} are consistent within academic content areas but not across all STAAR assessments.

For all STAAR science and social studies, STAAR EOC mathematics, and STAAR grades 4 and 7 writing assessments (as is evident in Tables 9.2 and 9.3 in Chapter 9), a scale score of 4000 represents the recommended Level II performance standard. In addition, those scales' standard deviations were set to 500. These values can be substituted into Equation (4) to provide a scaling equation specific to STAAR science and social studies, STAAR EOC mathematics, and STAAR grades 4 and 7 writing assessments.

$$SC_\theta = \frac{500}{\sigma_\theta} \times \theta + \left[4000 - \frac{500}{\sigma_\theta} \times \theta_{Level\_II} \right] \quad (5)$$

Thus, for the STAAR EOC mathematics, science, and social studies content areas, the cumulative score target is 3 x 4000 = 12,000.

For all English STAAR EOC assessments, a scale score of 2000 represents the recommended Level II performance standard. This scale score value is half as large as corresponding scale score cuts in other content areas because there are six English STAAR EOC assessments. Thus, the cumulative score target in English is identical to other content areas: 6 x 2000 = 12,000. Lastly, the EOC English scales' standard deviations were set to 250. These values can be substituted into Equation (4) to provide a scaling equation specific to STAAR EOC English assessments.

$$SC_\theta = \frac{250}{\sigma_\theta} \times \theta + \left[2000 - \frac{250}{\sigma_\theta} \times \theta_{Level\_II} \right] \quad (6)$$

It is important to note that although Level II scale score cuts are fixed across horizontally scaled assessments within content areas, Level III cuts vary across all STAAR assessments. These Level III cuts do not, however, vary over time. The fixed scale scores to be associated with Level III cuts (both phase- in and recommended) were calculated by substituting Level-III-specific θ values into Equations (5) and (6). For STAAR science and social studies, STAAR EOC mathematics, and STAAR grades 4 and 7 writing assessments, the following equation was used:

$$SC_\theta = \frac{500}{\sigma_\theta} \times \theta_{Level\_III} + \left[4000 - \frac{500}{\sigma_\theta} \times \theta_{Level\_II} \right] \quad (7)$$

For English STAAR EOC assessments, the following equation was used:

$$SC_\theta = \frac{250}{\sigma_\theta} \times \theta_{Level\_III} + \left[2000 - \frac{250}{\sigma_\theta} \times \theta_{Level\_II} \right] \quad (8)$$

In Equations (7) and (8), θ_{Level_III} refers to the Rasch-based performance estimates specific to Level III (phase-in and recommended) performance standards.

VERTICAL SCALING CONSTANTS

Similar to the horizontally scaled STAAR assessments, vertically scaled STAAR scale scores represent linear transformations of Rasch-based performance estimates (θ). Vertically scaled

scores, however, also include an extra scaling constant (V_g) that varies across each grade (g). This is given by the equation below:

$$SC_\theta = A \times (\theta - V_g) + B \tag{9}$$

The scaling constants *A* and *B* in Equation (9) are derived in the same way for the vertically scaled assessments, except that the scale score at the Level II cut is fixed for only the final assessment in the vertical scale (STAAR grade 8 reading and mathematics; STAAR Spanish grade 5 reading) and the standard deviation is taken across all of the assessments. The *A* scaling constant is calculated as follows:

$$A = \frac{\sigma_{sc}}{\sigma_\theta} \tag{10}$$

In Equation (10), σ_{sc} represents the desired standard deviation of the scale across all assessments, while σ_θ represents the standard deviation of Rasch-based θ values among a sample group. For the STAAR 3–8 vertical scales, the sample group consisted of all students who took the assessment across the vertical scale in spring 2012.

The *B* scaling constant is calculated as follows:

$$B = SC_{Level\_II} - \frac{\sigma_{sc}}{\sigma_\theta} \times \theta_{Level\_II} \tag{11}$$

In Equation (11), SC_{Level_II} represents the desired scale score at the Level II cut for the final assessment in the vertical scale, and θ_{Level_II} represents the approved Level II performance standard (in Rasch units) for the final assessment in the vertical scale. As in Equation (10), σ_{sc} represents the desired standard deviation of the scale, while σ_θ represents the standard deviation of Rasch-based θ values in the sample group. Using Equation (9) and substituting Equation (10) for *A* and Equation (11) for *B*, the full STAAR vertical scaling equation is shown below.

$$SC_\theta = \frac{\sigma_{sc}}{\sigma_\theta} \times (\theta - V_g) + \left[SC_{Level\_II} - \frac{\sigma_{sc}}{\sigma_\theta} \times \theta_{Level\_II} \right] \tag{12}$$

For the STAAR 3–8 mathematics and English reading vertical scales (as is evident in Table 9.3 in Chapter 9), a scale score of 1700 represents the recommended Level II performance standard for the grade 8 assessment. In addition, those scales' standard deviations were set to 150. These values can be substituted into Equation (12) to provide a scaling equation specific to the mathematics and English reading vertical scaled assessments.

$$SC_\theta = \frac{150}{\sigma_\theta} \times (\theta - V_g) + \left[1700 - \frac{150}{\sigma_\theta} \times \theta_{Level\_II} \right] \quad (13)$$

For the STAAR Spanish grade 5 reading assessment, a scale score of 1582 represents the recommended Level II performance standard. This scale score is set to the equivalent value as the Level II performance standard on the STAAR English grade 5 reading assessment. The Spanish reading vertical scale's standard deviations was also set to 150. These values can be substituted into Equation (12) to provide a scaling equation specific to STAAR Spanish grades 3–5 reading vertical scale.

$$SC_\theta = \frac{150}{\sigma_\theta} \times (\theta - V_g) + \left[1582 - \frac{150}{\sigma_\theta} \times \theta_{Level\_II} \right] \quad (14)$$

It is important to note that although Level II scale score cuts is fixed for the highest grade in the vertical scale, the Level II cuts for the other assessments in the vertical scale will vary. These Level II cuts, as well as the Level III cuts, do not vary over time. The fixed scale scores to be associated with the lower grades' Level II cuts (both phase- in and recommended), and all Level III cuts were calculated by substituting Level II and Level III-specific θ values into Equations (13) and (14) for each grade.

Figures 10.1–10.3 illustrate the Level II and Level III cut scores for mathematics, English reading, and Spanish reading, respectively. The STAAR vertical scales have the following characteristics:

- They range from approximately 600 to 2300 scale score points.
- The Level II cut score is 1700 for grade 8 mathematics and reading.
- The Level II cut score is 1582 for STAAR Spanish grade 5 reading, which is the same for STAAR English grade 5 reading.
- Level II cut scores increase across grades within a content area.
- Level III cut scores increase across grades within a content area.

Figure 10.1 STAAR 3–8 Mathematics Final Recommended Cut Scores

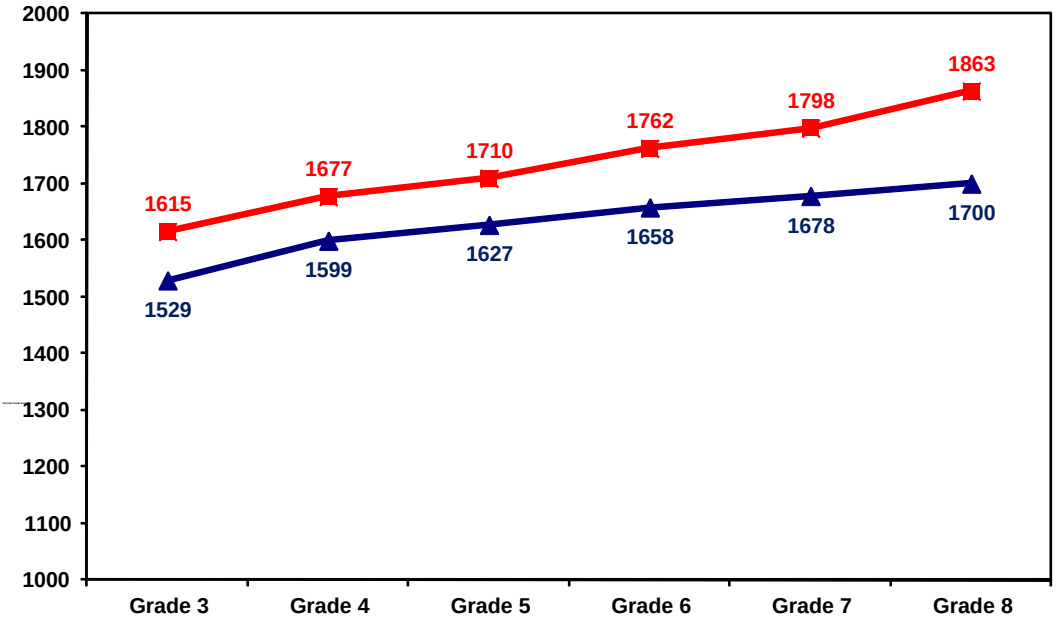


Figure 10.2 STAAR English 3–8 Reading Final Recommended Cut Scores

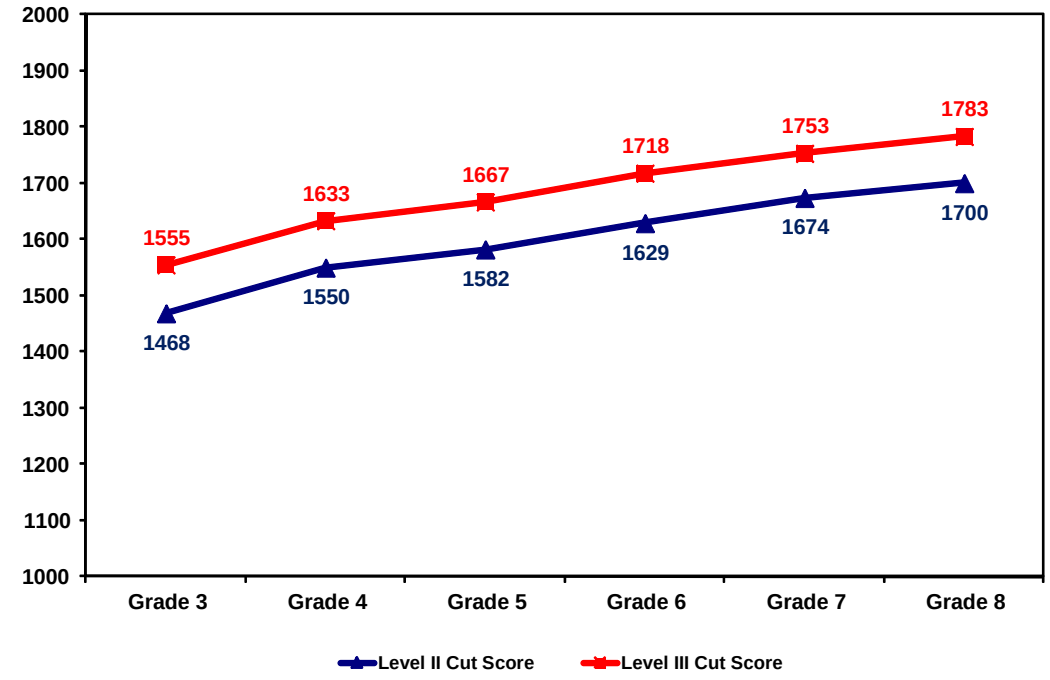
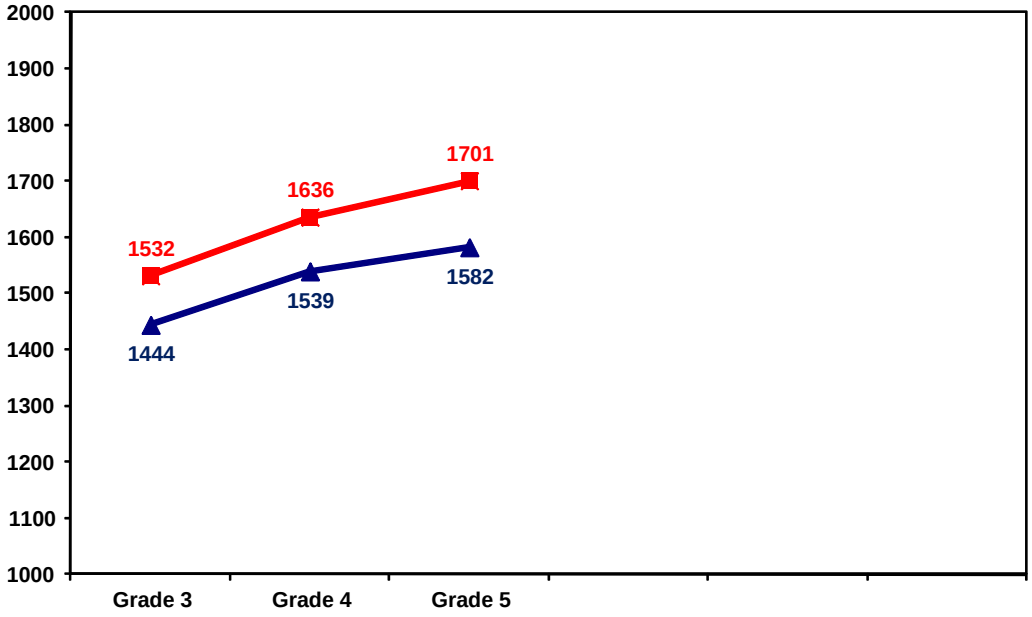


Figure 10.3 STAAR Spanish 3–5 Reading Final Recommended Cut Scores



Spring 2012 Scale Scores

For horizontal scales, using Equations (5)–(8), scale score tables were constructed for each of the 2012 STAAR assessments. It should be noted that across STAAR assessments, scale score ranges are truncated at 0 (that is, there are no negative scale scores) and are not truncated at the top of the scale. Although spring 2012 STAAR scales have an overall minimum of 0 and an overall maximum of 7480, roughly 99% of students' scale scores in science, social studies, EOC mathematics, and grades 4 and 7 writing fall between 2000 and 6000. Likewise, roughly 99% of students' scale scores in EOC English fall between 1000 and 3000.

For vertical scales, using Equations (13)–(14), scale-score tables were constructed for each of the 2012 STAAR assessments. The scale scores across all the vertically scaled assessments for the spring 2012 administration range from a minimum of 631 to a maximum of 2233. Ninety-nine percent of these vertical scale scores for spring 2012, however, fall between the range of 1000 to 2000.

The spring 2012 STAAR score tables can be found on the TEA website: http://www.tea.state.tx.us/student.assessment/staar/convtables/yr12/.

Rounding Rules

Although the scale score systems described above were established to facilitate approved Level II performance standards (in Rasch units) that always correspond to specific values (e.g., 2000 on English tests and 4000 on other horizontally scaled assessments), there is no guarantee that those Level II Rasch values will be obtainable on a given test form. The same is true for minimum scores—established one CSEM below Level II performance standards, as described in Chapter 8. The following sections describe rounding rules established to accommodate the various test forms required for implementation of the STAAR program

PERFORMANCE STANDARDS

The scale scores associated with Level II phase-in 1, Level II phase-in 2, and Level II recommended always appear in published scale-score tables. Similarly, phase-in and recommended Level III scale-score cuts appear in all score tables. Because the Rasch-based ability estimates associated with each obtainable raw score vary across administrations, those scale scores may not appear without use of some rounding procedure. Therefore, the calculated scale scores closest to those listed above are rounded to their target values when appropriate. For example, for English assessments, the calculated scale score closest to 1875 is rounded to 1875, the scale score closest to 1950 is rounded to 1950, and so on. The same rule applies to phase-in and recommended Level III standards, which vary across assessments.

MINIMUM SCORES

None of the calculated scale scores in published score tables is rounded to match established minimum scores; students must simply attain a scale score at or above established minimums (presented in Table 9.2 in Chapter 9) in order for that score to count toward the cumulative score requirement. However, CSEMs in scale-score units were required to establish the STAAR EOC program's minimum scores, which fall one CSEM below applicable Level II scale-score cuts. To convert a CSEM to scale score units, the *A* constant from Equation (2) is required.

$$CSEM_{SC} = A \times CSEM_{\theta} \tag{15}$$

In Equation (15), $CSEM_{sc}$ represents the CSEM in scale score units, while $CSEM_{sc}$ represents the CSEM in Rasch units. The CSEM is test-form dependent, so spring 2012 score tables were used to calculate appropriate CSEMs and establish minimum scores. The process involved three steps. An illustrative example of these steps to establish the final recommended minimum score for the spring 2012 STAAR Algebra I assessment is shown below.

1. Locate the calculated scale score closest to 4000 and round that scale score to 4000. For Algebra I, this scale score corresponds to a raw score of 34.
2. Convert the CSEM at that score from Rasch units (0.290) to scale score units, applying Equation (9):`

$$CSEM_{SC} = A \times CSEM_{\theta}$$
$$128 = 441.1057 \times 0.290$$

3. Subtract the CSEM in scale score units (128) from the scale score cut (4000) to arrive at the minimum score in scale score units (4000 – 128 = **3872**).

Minimum scores for all phase-in and recommended standards across STAAR EOC assessments are noted in the score tables on the TEA website: http://www.tea.state.tx.us/student.assessment/staar/convtables/yr12/.

Chapter 11: Review of Performance Standards

This chapter provides details about Step 9 of the nine-step STAAR standard-setting process, which is focused on reviewing performance standards. The sections in this chapter include

- Legislative Requirement
- STAAR Standards Review Plan

Legislative Requirement

State statute requires that performance standards for the STAAR program be reviewed at least once every three years. The specific legislation in the Texas Education Code is as follows.

Section 39.0242

(d) The agency shall continue to gather data and perform studies as provided under this section at least once every three years. If the data do not support the correlation between student performance standards and college readiness, the Commissioner of Education, in collaboration with the Commissioner of Higher Education, shall revise the standard of performance considered to be satisfactory.

(e) Based on the data collected and studies performed periodically under Subsection (d), the Commissioner shall increase the rigor of the performance standard established under Section 39.0241(a) as the Commissioner determines necessary.

Section 39.0241(a) requires the establishment of satisfactory performance standards on the STAAR assessments. Refer to Appendix 1 for more details.

STAAR Standard Review Plan

Initial STAAR performance standards were established in 2012. To be in compliance with state statute, the performance standards must be reviewed by 2015. The current plan is to review the STAAR performance standards in fall 2014. Doing so will allow for the cut scores to be considered based on operational and motivated data for all the STAAR assessments.

Before the planned fall 2014 standards review, additional empirical studies will be conducted based on student test data collected from the 2011–2012, 2012–2013, and 2013–2014 school years. This longitudinal data can be used to evaluate the alignment of the performance standards within STAAR 3–8 and as students progress from middle school to high school. There are also plans to collect student test data from additional external assessments, such as the Advanced Placement (AP), International Baccalaureate (IB), Preliminary SAT (PSAT) or PLAN tests, or related postsecondary readiness measures, such as indicators of workforce or military readiness. A process similar to the one used in the initial standard-setting process will be followed to summarize, evaluate, and incorporate the empirical studies into the standards review process.

In subsequent reviews of the performance standards (for example, before the 2017–2018 school year), longitudinal data collected by following Texas students from high school into college may also be used to evaluate the reasonableness of the postsecondary-readiness standards during the standards review process.

As seen in the STAAR standard-setting process described in this report, unlike in previous Texas assessment programs, the legislative intent is for the process of setting performance standards in STAAR to no longer be a one-time event with a single committee of educators and content experts. Instead, it will be an ongoing iterative process that involves different types of stakeholders, is informed by empirical studies, and results in a comprehensive system that links the performance standards on previous and successive assessments in the same content area.

Appendix 1: State Statutes on STAAR Performance Standards

TEXAS EDUCATION CODE SECTION 28.021

(a) Except as provided by Subsection (b) or (e), a student may not be promoted to:

(1) the sixth grade program to which the student would otherwise be assigned if the student does not perform satisfactorily on the fifth grade mathematics and reading assessment instruments under Section 39.023; or

(2) the ninth grade program to which the student would otherwise be assigned if the student does not perform satisfactorily on the eighth grade mathematics and reading assessment instruments under Section 39.023.

TEXAS EDUCATION CODE SECTION 39.0233

(b) In addition to the questions adopted under Subsection (a), the agency shall adopt a series of questions to be included in an end-of-course assessment instrument administered under Section 39.023(c) to be used for purposes of identifying students who are likely to succeed in an advanced high school course. A school district shall notify a student who performs at a high level on the questions adopted under this subsection and the student's parent or guardian of the student's performance and potential to succeed in an advanced high school course.

TEXAS EDUCATION CODE SECTION 39.024

(c) Before the beginning of the 2011-2012 school year, the agency, in collaboration with the Texas Higher Education Coordinating Board, shall gather data and conduct research studies to substantiate the correlation between a certain level of performance by students on the Algebra II and English III end-of-course assessment instruments and college readiness.

(e) Based on the results of the studies conducted under Subsection (c), the commissioner of education and the commissioner of higher education shall establish student performance standards for the Algebra II and English III end-of-course assessment instruments indicating that students have attained college readiness.

(f) The agency, in collaboration with the Texas Higher Education Coordinating Board, shall conduct research studies similar to the studies conducted under Subsection (c) for the appropriate science and social studies end-of-course assessment instruments. If the commissioner of education, in collaboration with the commissioner of higher education, determines that the research studies conducted under this subsection substantiate a correlation between a certain level of performance by students on science and social studies end-of-course assessment instruments and college readiness, the commissioner of education, in collaboration with the commissioner of higher education, as soon as practicable, may establish student performance standards for the science and social studies end-of-course assessment instruments indicating that students have attained college readiness.

TEXAS EDUCATION CODE SECTION 39.0241

(a) The commissioner shall determine the level of performance considered to be satisfactory on the assessment instruments.

(a-2) For the purpose of establishing performance across grade levels, the commissioner shall establish:

(2) the performance standards for the Algebra I and English II end-of-course assessment instruments, as determined based on studies under Section 39.0242 that correlate student performance on the Algebra I and English II end-of-course assessment instruments with *student performance on the Algebra II and English III assessment instruments*

(3) the performance standards for the English I end-of-course assessment instrument, as determined based on studies under Section 39.0242 that correlate student performance on the English I end-of-course assessment instrument with *student performance on the English II assessment instrument*

(4) the performance standards for the grade eight assessment instruments, as determined based on studies under Section 39.0242 that correlate student performance on the grade eight assessment instruments with student performance on the Algebra I and English I end-of-course assessment instruments in the same content area; and

(5) the performance standards on the assessment instruments in each of grades three through seven, as determined based on studies under Section 39.0242 that correlate student performance in the same content area on the assessment instrument for each grade with student performance on the assessment instrument in the succeeding grade.

TEXAS EDUCATION CODE SEC. 39.0242

(a) During the 2009-2010 and 2010-2011 school years, the agency shall collect data through:

(1) the annual administration of assessment instruments required under Section 39.023(a) in grades three through eight; and

(2) the administration to a sufficiently large sample of students throughout the state of end-of-course assessment instruments required under Section 39.023(c) for the purpose of setting performance standards.

(b) Before the beginning of the 2011-2012 school year, the agency shall analyze the data collected under Subsection (a) to substantiate:

(1) the correlation between satisfactory student performance for each performance standard under Section 39.0241 on the grade three, four, five, six, or seven assessment instruments with satisfactory performance under the same performance standard on the assessment instruments in the same content area for the next grade level;

(2) the correlation between satisfactory student performance for each performance standard under Section 39.0241 on the grade eight assessment instruments with satisfactory performance under the same performance standard on the Algebra I and English I end-of-course assessment instruments in the same content area;

(3) the correlation between satisfactory student performance for each performance standard under Section 39.0241 on the English I end-of-course assessment instrument with satisfactory performance under the same performance standard on the English II end-of-course assessment instrument;
(4) the correlation between satisfactory student performance for each performance standard under Section 39.0241 on the English II end-of-course assessment instrument with satisfactory performance under the same performance standard on the English III end-of-course assessment instrument; and
(5) the correlation between satisfactory student performance for each performance standard under Section 39.0241 on the Algebra I end-of-course assessment instrument with satisfactory performance under the same performance standard on the Algebra II end-of-course assessment instrument.

(c) Studies under this section must include an evaluation of any need for remediation courses to facilitate college readiness.
(d) The agency shall continue to gather data and perform studies as provided under this section at least once every three years. If the data do not support the correlation between student performance standards and college readiness, the commissioner of education, in collaboration with the commissioner of higher education, shall revise the standard of performance considered to be satisfactory.
(e) Based on the data collected and studies performed periodically under Subsection (d), the commissioner shall increase the rigor of the performance standard established under Section 39.0241(a) as the commissioner determines necessary.

TEXAS EDUCATION CODE SECTION 39.025

(a) The commissioner shall adopt rules requiring a student participating in the recommended or advanced high school program to be administered each end-of-course assessment instrument listed in Section 39.023(c) and requiring a student participating in the minimum high school program to be administered an end-of-course assessment instrument listed in Section 39.023(c) only for a course in which the student is enrolled and for which an end-of- course assessment instrument is administered. A student is required to achieve, in each subject in the foundation curriculum under Section 28.002(a)(1), a cumulative score that is at least equal to the product of the number of end-of-course assessment instruments administered to the student in that subject and a scale score that indicates satisfactory performance, as determined by the commissioner under Section 39.0241(a). A student must achieve a minimum score as determined by the commissioner to be within a reasonable range of the scale score under Section 39.0241(a) on an end-of-course assessment instrument for the score to count towards the student's cumulative score.

TEXAS EDUCATION CODE SECTION 39.036

(a) The agency shall develop a vertical scale for assessing student performance on assessment instruments administered under Sections 39.023(a)(1) and (2) in a manner that allows the agency to compare the performance of a student on the assessment instruments from one grade level to the next.

(b) The commissioner shall adopt rules necessary to implement this section.

Appendix 2: Empirical Studies Methodological Notes

Description and Purpose of Empirical Studies

Legislative requirements and the availability of data resulted in different empirical studies for the high school assessments compared to the elementary and middle school assessments. Tables A2.1 and A2.2 give the description and purpose of each empirical study that was conducted to inform the initial STAAR EOC and STAAR 3–8 standard-setting processes, respectively.

Table A2.1: Empirical Studies for the STAAR EOC Standard-Setting Process

STAAR EOC Empirical Study	Description and Purpose of Study
Empirical links between courses (*Linking studies*)	Studies empirically linked student performance on STAAR EOC assessments in the same content area (i.e., mathematics or English). The results of the studies were 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 (*Bridge studies*)	Studies compared 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 (*Grade correlation studies*)	Studies compared 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 (*External validity studies*)	Studies established empirical links between student performance on the STAAR EOC assessments with that on the SAT and ACT tests. The results of the studies externally validated the STAAR performance standards with tests taken nationally for the purpose of college admissions.

Table A2.1 cont.: Empirical Studies for the STAAR EOC Standard-Setting Process

STAAR EOC Empirical Study	Description and Purpose of Study
Comparison with THEA and ACCUPLACER (*External validity studies*)	Studies established empirical links between student performance on the STAAR EOC assessments with that on the ACCUPLACER and THEA tests. The results of the studies externally validated the STAAR performance standards with tests taken for the purpose of college placement.
Comparison with NAEP and PISA	Studies examined 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 were used to evaluate the rigor of the STAAR performance standards relative to performance standards established for national and international assessment instruments.
College students taking STAAR Algebra II and English III	Studies compared STAAR performance of college students who were successful in an entry-level college course to those who were not successful. These studies provided a direct measure of college student performance on the STAAR EOC assessments.

Table A2.2: Empirical Studies for the STAAR 3–8 Standard-Setting Process

STAAR 3–8 Empirical Study	Description and Purpose of Study
Empirical links with EOC (*Linking studies*)	Studies empirically linked student performance on STAAR grade 8 assessments and the grade 7 writing assessment with EOC assessments in the same content area (i.e., mathematics or reading). The results of the studies were used to inform the alignment of performance standards between middle school and high school 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.

Table A2.2 cont.: Empirical Studies for the STAAR 3–8 Standard-Setting Process

STAAR 3–8 Empirical Study	Description and Purpose of Study
Empirical links across grades (*Linking studies*)	Studies empirically linked student performance on STAAR assessments for grades 3–8 in the same content area. The results of the studies were 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 assessment in the same content area.
Vertical Scale	Under TEC §39.036, TEA is required to develop a vertical scale for assessing student performance in grades 3–8 for reading and mathematics. Vertical scales allow the comparison of student performance across grades within a content area. The results of the studies were used to inform the alignment of performance standards across assessments.
Comparison with TAKS (*Bridge studies*)	Studies compared performance on STAAR assessments with corresponding TAKS assessments to ensure that the performance standards for STAAR are more rigorous than TAKS performance standards.
Comparison with ReadiStep and EXPLORE (*External validity studies*)	Studies established empirical links between student performance on the STAAR grade 8 assessments and the grade 7 writing assessment with that on the ReadiStep and EXPLORE tests, which are linked to SAT and ACT, respectively. The results of the studies externally validated the STAAR performance standards with tests taken nationally for the purpose of determining if students are on track for college readiness.
Comparison with NAEP	Studies examined impact data from the National Assessments of Educational Progress (NAEP). The results of these studies were used to evaluate the rigor of the STAAR performance standards relative to performance standards established for a national assessment instrument.

Statistical Methods: Equipercentile Linking

The specific steps for the equipercentile equating between a TAKS assessment (*X*) and a STAAR assessment (*Y*) include:

1. Let $T_X \in (1000, 3000)$ (note: the upper and lower boundary will be determined by the observed TAKS scores, or, if applied, pre-smoothing) with an interval of 1 between each T_X on the TAKS assessment (*X*). For each T_X, the following equipercentile function was used to find the corresponding raw score on the given STAAR assessment (*Y*):

$$\frac{\Pr(X < T_X) + 0.5 * \Pr(X = T_X) - \Pr\left(Y < u * (T_X)\right)}{\Pr\left(Y = u * (T_X)\right)} + u * (T_X) - 0.5$$

   Where $u^*(T_X)$ is smallest raw score on the STAAR assessment such that,

$$\Pr(X < T_X) + 0.5 * \Pr(X = T_X) < \Pr\left(Y \leq u * (T_X)\right)$$

2. The resulting table will list the STAAR raw score that has the same percentile rank (Pr) as each TAKS score. Under this approach, many consecutive TAKS scale scores may be expected to correspond with the same STAAR raw score (e.g., the TAKS scale scores 2073 through 2103 may all correspond with a STAAR raw score of 19). As such, any given STAAR raw score may be associated with a band of TAKS scale scores (e.g., 19 = 2073-2103). The equipercentile equating results were summarized using this STAAR raw score-to-TAKS score band approach.
3. The STAAR raw score associated with the TAKS score band that includes the *Met Standard* scale score cut (i.e., 2100) served as the result for STAAR assessments.
4. When a single STAAR form is administered to a population of examinees, a single table is produced providing STAAR scores on the Rasch scale (θ) that correspond to each obtainable STAAR raw score. This "raw-score-to-theta" table facilitated the interpretation of the study results using either raw scores or θ estimates. Because a one-to-one correspondence exists between raw scores and θ estimates, the equipercentile equating function presented in Step 1 can use either measure; the study results will be equivalent.

Table A2.3 provides an example of the STAAR to TAKS comparison study results for STAAR Algebra I to TAKS grade 9 mathematics. In Table A2.3, the TAKS score bands for grade 9 mathematics are mapped to obtainable raw scores on the STAAR Algebra I 2012 base test. In this example, the TAKS grade 9 mathematics scale score needed to attain *Met Standard* was a value of 2100. An equipercentile linking procedure between STAAR Algebra I raw scores and TAKS scale scores indicated that a TAKS score band of 2073-2103 (which includes 2100) corresponded to an Algebra I raw score of 19. This raw score corresponded to an Algebra I proficiency estimate (on the Rasch scale) of roughly -0.15. Therefore, based on this study, a score of 19 on the Algebra I test (or, equivalently, a θ estimate of -0.15) is associated with the TAKS *Met Standard* cut. The Algebra I raw scores and θ values and the TAKS mathematics scores in Table A2.3 do not represent study results; they are provided for illustrative purposes.

Table A2.3: Example Comparison Study for STAAR Algebra I to TAKS Grade 9 Mathematics

STAAR Algebra I Raw Score	TAKS Grade 9 Mathematics Scale Score	STAAR Algebra I Proficiency Estimate (θ)
…	…	…
17	2026 - 2056	-0.33
18	2057 - 2072	-0.24
19	**2073 - 2103**	**-0.15**
20	2104 - 2118	-0.07
21	2119 - 2135	0.02
…	…	…

Statistical Methods: OLS Regression

In each regression model, a given external test score (*Y*) was specified as the dependent variable, and STAAR proficiency estimates (θ) were specified as the independent variable. The simple linear regression model is specified as follows:

$$Y_i = \beta_0 + \beta_1 STAAR\theta_i + \varepsilon_i$$

The β_0 and β_1 (intercept and slope) parameters are estimated via OLS regression. The resulting parameter estimates ($\widehat{\beta_0}$ and $\widehat{\beta_1}$) can then be used to compute projected *Y* score values ($\widehat{Y_i}$) conditional on θ values:

$$\widehat{Y_i} = \widehat{\beta_0} + \widehat{\beta_1} STAAR\theta_i$$

A projected *Y* score was calculated for each θ value between -3 and 3, inclusive, at intervals of 0.1 (i.e., -3, -2.9, -2.8, …, 3). Each projected score was rounded to the nearest integer. Predicted scores were summarized in a projection table. An example is provided in Table A2.4. SAT scores and θ values in Table A2.4 do not represent study results; they are provided for illustrative purposes.

Table A2.4: Example Projection Table (STAAR EOC → SAT)

STAAR EOC English III Reading Proficiency Estimate (θ)	SAT Critical Reading Score
…	…
-1.5	358
-1.4	370
-1.3	392
-1.2	418
-1.0	438
…	…

Statistical Methods: Logistic Regression

In this approach, dichotomized external scores (*Y*) were specified as the dependent variable, and STAAR proficiency estimates are specified as the independent variable. More specifically, for each external score point *k*, the dichotomous variable, Y_i was constructed for each student (*i*) such that

- Y_i = 0, if the student's external score is less than *k*;
- *Y* = 1, if the student's external score is greater than or equal to *k* .

For example, a dichotomous variable for the SAT score of 500 was created such that students who scored below 500 would receive a 0, and students who scored at or above 500 would receive a 1.

Let STAARθ_i represent the STAAR proficiency estimate for each student (*i*) and *p* be the probability that Y_i = 1, given STAARθ_i; that is, *p* = Prob (Y_i = 1 | STAARθ_i). Using logistic regression, the regression coefficients (β_0 and β_1) in the equation below were obtained:

$$logit(p)_i = \beta_0 + \beta_1 STAAR\theta_i + \varepsilon_i$$

Where $\hat{p}_i$ is the expected probability that Y_i = 1 and $logit(p)_i = ln\frac{p}{1-p}$.

For each STAAR θ value between -3 and 3, inclusive, at intervals of 0.1, the expected probability ($\hat{p}$) for getting score *k* or higher on the external test was computed. The probability values can be computed by substituting each θ value into the following equation:

$$\hat{p}_i = \frac{\exp\left(\widehat{\beta_0} + \widehat{\beta_1} STAAR\theta_i\right)}{1 + \exp\left(\widehat{\beta_0} + \widehat{\beta_1} STAAR\theta_i\right)}$$

These calculations yielded one probability corresponding to each combination of θ (-3, -2.9, -2.8, …, 3) and external score (e.g., SAT score at 200, 210, 220, …, 800). Those probabilities were displayed in an expectancy table. An example is provided in Table A2.5. In this scenario, a student with an English III reading θ = -1.4 has a 53% chance of obtaining at least a 16 on the ACT reading assessment. ACT score probabilities and θ values in Table A2.5 do not represent study results; they are provided for illustrative purposes.

Table A2.5: Example Expectancy Table (STAAR EOC → ACT)

STAAR EOC English III Reading Proficiency Estimate (θ)	Probability of an ACT Reading Score of At Least…				
	14	15	16	17	18
…					
-1.6	54	48	43	37	32
-1.5	58	53	48	42	37
-1.4	63	58	53	47	42
-1.3	68	63	58	52	47
-1.2	72	68	63	57	52
…					

Statistical Methods: Item Response Theory

The TAKS and STAAR assessments are scaled and equated using an item response theory model known as the Rasch Partial-Credit Model (RPCM) to place test items and measures of student proficiency on the same scale across assessments. The RPCM is an extension of the Rasch one-parameter Item Response Theory model attributed to Georg Rasch (1966), as extended by Wright and Stone (1979), Masters (1982), Wright and Masters (1982), and Linacre (2001). The RPCM maintains a one-to-one relationship between scale scores and raw scores, meaning each raw score is associated with a unique scale score.

The RPCM is defined by the following mathematical measurement model where, for a given item involving *m* + 1 score categories, the probability of person *n* scoring *x* on prompt *i* is given by:

$$P_{xni} = \frac{\exp \sum_{j=0}^{x} (\theta_n - \delta_{ij})}{\sum_{k=0}^{m_i} \exp \sum_{j=0}^{x} (\theta_n - \delta_{ij})}, \qquad x = 0, 1, \ldots, m_i$$

The RPCM provides the probability of a student scoring *x* on the *m* steps of question/prompt, *i* as a function of the student's proficiency level, θ_n, and the step difficulties, δ_{ij}, of the *m* steps in prompt *i*. (Refer to Masters, 1982, for an example.) Note that for multiple-choice and gridded-response questions, there are only two score categories: (a) 0 for an incorrect response and (b) 1 for a correct response, in which case the RPCM reduces to the standard Rasch one-parameter IRT model, and the resulting single-step difficulty is more properly referred to as an item difficulty. The underlying Rasch scale enables the maintenance of equivalent performance standards across test forms and places all items on a common Rasch scale. RPCM was used for two of the empirical studies: the STAAR-to-TAKS comparison study and the linking study based on the vertical scale analyses.

STAAR-TO-TAKS COMPARISON STUDIES

For most of the STAAR-to-TAKS comparison studies in grades 3–8, STAAR items were embedded in the spring 2011 TAKS administration. The RPCM resulted in the STAAR and TAKS items and proficiency levels being placed on the same Rasch scale. The TAKS *Met Standard* proficiency level, θ_{ms} , was identified by finding the STAAR proficiency level that equaled the TAKS *Met Standard* proficiency level. An example is provided in Table A2.6. The TAKS *Met Standard* proficiency level and θ values in Table A2.6 do not represent study results; they are provided for illustrative purposes.

Table A2.6: Example STAAR-to-TAKS Comparison

STAAR Grade 8 Reading Proficiency Estimate (θ)	TAKS Grade 8 Performance Standard
…	…
-0.5	
-0.4	
-0.3	TAKS Met Standard
-0.2	
-0.1	
…	…

This analysis method was one of two methodologies in the STAAR-to-TAKS comparison study results. Additional information can be found on the TEA website at www.tea.state.tx.us/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=2147504930&libID=2147504925

VERTICAL SCALE STUDIES

STAAR items from adjacent grade levels were embedded in spring 2012 STAAR tests (e.g., grade 4 mathematics items were included on grade 5 mathematics test forms). The items were embedded in field-test positions and did not count toward student scores. The RPCM placed all STAAR items across grades within a content area on a common scale. The vertical scale studies determined the difference in test difficulty between adjacent grades for STAAR 3–8 reading and mathematics assessments.

A two-step calibration procedure was used. The first calibration step calibrated only the base-test items within a grade level together using all available student data. The second calibration step calibrated for each grade level the base-test items, on-grade level vertical linking items, and off-grade level vertical linking items through an incomplete data matrix (IDM). Equating constants were then determined by finding the difference between the means of the base-test items $(\bar{b}_{BTonly})$ from the first calibration and the base-test items $(\bar{b}_{BTandVS})$ from the second calibration and adding the equating constant to all the vertical linking items.

$$C^{*} = \bar{b}_{BTonly} - \bar{b}_{BTandVS}$$

This difference (C^*) is the equating constant that is added to all the vertical linking items.

Once all the items within a grade level were on the same scale, adjacent grade vertical scaling constants were calculated by comparing the mean item difficulties for the vertical linking items across adjacent grades (mean/mean method) using the combined vertical linking item set (both lower and upper grade level vertical linking items).

Rasch Mean Square infit and residual analyses were used to remove vertical linking items with anomalous results, as long as removing the items did not result in poor content representation. Additionally, outlier items were removed based on studentized residuals and visual inspections of the fit plots.

After finding the vertical linking constants between adjacent grades, cumulative linking constants were defined from the base grade to any grade levels that were not adjacent to the base grade. The base grade level was grade 8 for STAAR English reading and mathematics and grade 5 for STAAR Spanish reading. The vertical scaling constant at those grades was set at zero, and the vertical scaling constant at the other grades was calculated based on that end point. Figure A2.1 illustrates an example of vertical scaling constants from grades 3–8.

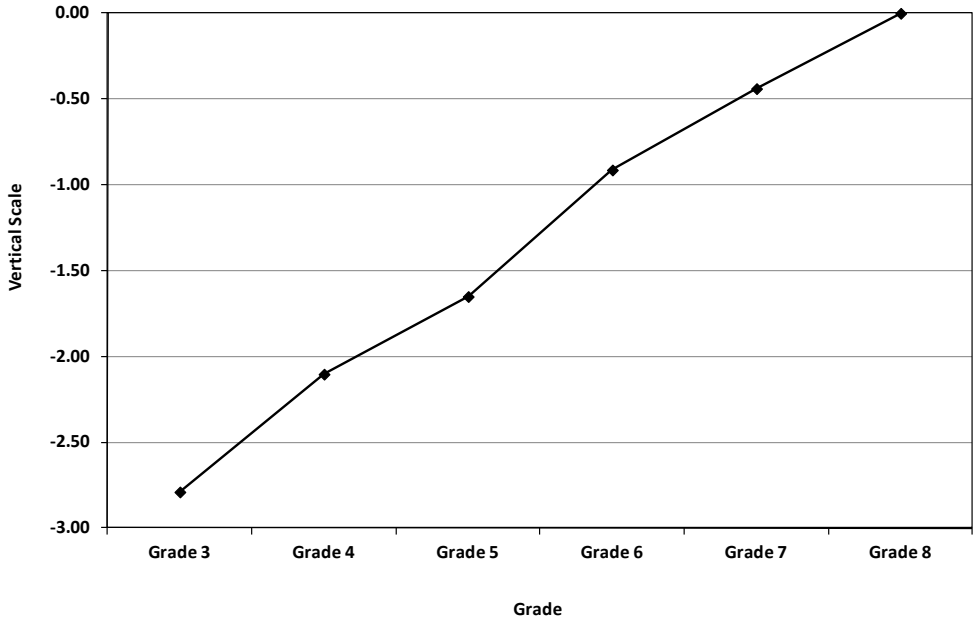


Figure A2.1: Example Vertical Scale Constants

Appendix 3: STAAR EOC Empirical Studies Quality Summary

Study Name	Motivation	Representativeness	Sample Size	Correlation	Content Overlap	Overall
Algebra II – ACT Mathematics	★★★☆☆	★★★★☆	★★★★★	★★★☆☆	★★☆☆☆	★★★½☆
Algebra II – SAT Mathematics	★★★☆☆	★★★★☆	★★★★★	★★★☆☆	★★☆☆☆	★★★½☆
Algebra II – THEA Mathematics	★★★☆☆	★★★★☆	★★★☆☆	★★☆☆☆	★★☆☆☆	★★½☆☆
Algebra II – ACCUPLACER Algebra	★★★☆☆	★★★☆☆	★★★★☆	★★☆☆☆	★★☆☆☆	★★½☆☆
College Students Taking STAAR Algebra II	★★★☆☆	★★★☆☆	★★☆☆☆	★★★☆☆	N/A	★★★☆☆
English III Reading – ACT Reading	★☆☆☆☆	★★★☆☆	★★★★★	★★☆☆☆	★★★☆☆	★★½☆☆
English III Reading – SAT Critical Reading	★☆☆☆☆	★★☆☆☆	★★★★★	★★★☆☆	★★★★☆	★★★☆☆
English III Reading – THEA Reading	★☆☆☆☆	★★★★☆	★★★☆☆	★★☆☆☆	★★★☆☆	★★½☆☆
English III Reading – ACCUPLACER Reading	★☆☆☆☆	★★★★☆	★★★☆☆	★★★☆☆	★★☆☆☆	★★½☆☆
College Students Taking STAAR English III Reading	★★☆☆☆	★☆☆☆☆	★★☆☆☆	★☆☆☆☆	N/A	★½☆☆☆
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 Taking STAAR English III Writing	★★☆☆☆	★☆☆☆☆	★★☆☆☆	★☆☆☆☆	N/A	★½☆☆☆

Study Name	Motivation	Representativeness	Sample Size	Correlation	Content Overlap	Overall
Biology – ACT Science	★★★☆☆	★☆☆☆☆	★★★★★	★★★☆☆	★★☆☆☆	★★★☆☆
Biology – SAT Mathematics	★★★☆☆	★☆☆☆☆	★★★★★	★★★☆☆	☆☆☆☆☆	★★½☆☆
Chemistry – ACT Science	★★★☆☆	★★★☆☆	★★★★★	★★★☆☆	★★☆☆☆	★★★☆☆
Chemistry – SAT Mathematics	★★★☆☆	★★☆☆☆	★★★★★	★★★★☆	☆☆☆☆☆	★★★☆☆
Physics – ACT Science	★★★☆☆	★★★☆☆	★★★★★	★★★☆☆	★★☆☆☆	★★★☆☆
Physics – SAT Mathematics	★★★☆☆	★★★☆☆	★★★★★	★★★☆☆	☆☆☆☆☆	★★★☆☆
World Geography – ACT Reading	★★★☆☆	★★☆☆☆	★★★★★	★★★☆☆	☆☆☆☆☆	★★½☆☆
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		
★★☆☆☆	Data derive from stand-alone field tests and low-stakes operational STAAR administrations, are linked to motivated external assessments, and include constructed responses		
★★★☆☆	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		
★★★★☆	Data derive from some low-stakes and some high-stakes STAAR administrations, are linked to motivated external assessments, and may include constructed responses		
★★★★★	All data (STAAR assessments and external assessments) derive from high-stakes, motivated administrations		
	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 4: STAAR 3–8 Empirical Studies Summary

STAAR 3–8 to STAAR EOC

Studies empirically linked student performance on the STAAR grade 8 assessments and the grade 7 writing assessment with EOC assessments in the same content area (i.e., mathematics or English). The results of the studies were used to inform the alignment of performance standards between middle school and high school assessments.

For the STAAR grade 8 assessments and grade 7 writing, there were two potential data-collection designs available for linking to STAAR EOC: single-group design and coarsened exact matching. The following discusses the rationales for selecting the data-collection design for these studies. Both data sources were based on motivated high-stakes administrations. A single-group design was possible by using the spring 2011 TAKS data (with STAAR field-test items) to link to spring 2012 STAAR EOC assessments. This was possible because of the bridge study that previously established the relationship between TAKS and STAAR assessments in grades 3–8. The main drawback to this data collection design stemmed from the differences between the TAKS and STAAR assessments with respect to rigor and test blueprint.

The spring 2012 STAAR data for grade 8 assessments and grade 7 writing represented a second data source available to link to STAAR EOC. Although this did not allow for a single-group design, it did enable the use of motivated high-stakes data for the STAAR grade 8 assessments and the grade 7 writing assessment. Coarsened exact matching created matched samples for the two spring 2012 STAAR assessments using characteristics statistically associated with both tests. Those characteristics included demographic variables and the previous year's test scores. The quality of the linking study is contingent on whether the variables used to create the matched samples are highly correlated with the tests that are being linked. As a result, it is preferred to have the previous year's test scores used in the matching to be from the same content area as the tests being linked. For the STAAR grade 8 reading and mathematics assessments, both samples of students, took STAAR in 2012 and had previously taken TAKS grade 7 reading and mathematics assessments, respectively. For the STAAR grade 8 science and social studies assessments and the grade 7 writing assessment, the prior year's test scores in the same content area were not available to use as a matching variable. The only previous test scores were from the reading and mathematics content areas.

Both data-collection designs were evaluated for linking the STAAR grade 8 assessments and grade 7 writing assessment to STAAR EOC assessments. The correlations between the STAAR tests to be linked were examined for both methods. The availability of the prior year's test scores in the *same content area* was critical to creating the matching variables for coarsened exact matching and resulted in higher correlations than the single-group design. Assessments without the prior year's test scores in the same content area had lower correlations for the coarsened exact matching than compared to the single-group design. As a result, the coarsened exact matching was used for STAAR grade 8 reading and mathematics to EOC, and the single-group design was used for STAAR grade 8 science and social studies and the grade 7 writing

assessments. Table A4.1 shows the data-collection design, the sample sizes available for each analysis, and the correlations between the scores from the linked tests.

The representativeness of the assessments varied by linking study in terms of demographics, student proficiency, and sample size as a result of the availability of STAAR EOC data for different assessments. Courses that are often taken in ninth grade had large sample sizes given that the spring 2012 ninth graders were required to take STAAR as a graduation requirement. Courses that are generally taken at higher grade levels had lower sample sizes, which resulted in the representativeness of the linking samples being less similar to the population. Although the correlations in Table A4.1 indicated relatively strong relationships for each study, the studies with a low sample size were less generalizable. As a result, these studies were relied on less in determining the reasonable ranges, or "neighborhoods," and in terms of feedback during the standard-setting meetings. The linking studies based on coarsened exact matching had sample sizes almost double the student population, since student data were matched statistically across assessments to create pairs of scores for the analysis.

Table A4.1: 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

*Correlations are statistical measures of the relationships between scores on separate STAAR assessments. Correlations can range from -1 to 1; high positive values indicate strong positive relationships. For example, students with high STAAR grade 8 mathematics scores tend to have high STAAR Algebra I scores.

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 probability of attaining the final performance standard. These results informed the neighborhoods in which the standard-setting committees set cut scores for STAAR grade 8 assessments and grade 7 writing.

STAAR 3–8 Empirical Links Across Grades

Studies empirically linked student performance across grades within content areas for the STAAR 3–8 assessments (e.g., STAAR grade 5 reading to STAAR grade 6 reading). The results of the studies were used to inform the alignment of performance standards. This alignment should provide an advanced indicator about whether students are on track to meet the performance standards on a subsequent STAAR 3–8 assessment in the same content area.

The potential data collection designs for the across-grade linking studies varied because of the availability of data, similar to the options evaluated for STAAR grade 8 and grade 7 writing linking studies to STAAR EOC. Therefore, similar processes were followed in determining the data source. 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 avaible for linking. For reading and mathematics assessments, data were available for a single-group design using the spring 2011 TAKS data (with STAAR field-test items). Data were also available for most of the linking studies for coarsened exact matching, which requires an achievement composite in which the students to be matched have scores on a common prior assessment from the same content area. For these assessments, coarsened exact matching was used. The exceptions were the linking studies from grade 3 to grade 4 in which no common prior assessment was available. For example, grade 3 is the first grade tested in the STAAR program; therefore, no common prior assessment is 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 linking study from grade 4 writing to grade 7 writing and the linking study from grade 5 science to grade 8 science did not have a prior assessment available from the same content area. This was a similar issue for STAAR grade 8 and grade 7 to EOC, in which the prior year's assessments were in the reading and mathematics content areas. A single-group design was implemented. However, because of the expanse of years between administrations, there was not a single group of students available for the grade 4 writing to grade 7 writing study or the grade 5 science to grade 8 science. Therefore, the data-collection design was based on coarsened exact matching with prior year tests' scores from reading and mathematics content areas.

All data for the STAAR 3–8 links across grades were based on motivated high-stakes administrations. The representativeness of the assessments in terms of demographics and student proficiency was very similar to the student populations. Each linking study had large sample sizes. The linking studies based on coarsened exact matching had sample sizes almost double the student population, since student data were matched statistically across assessments to create pairs of scores for the analysis. Table A4.2 shows the sample sizes available for each analysis and the correlations between the scores from the linked tests.

Table A4.2: Sample Size and Correlation for STAAR 3–8 Empirical Links Across Grades

STAAR Assessment	STAAR Linked Test	Data Collection	Sample Size	Correlation*
Grade 3 mathematics	Grade 4 mathematics	Single-group design	304,339	0.70
Grade 4 mathematics	Grade 5 mathematics	Coarsened exact matching	505,393	0.67
Grade 5 mathematics	Grade 6 mathematics	Coarsened exact matching	568,104	0.75
Grade 6 mathematics	Grade 7 mathematics	Coarsened exact matching	583,174	0.76
Grade 7 mathematics	Grade 8 mathematics	Coarsened exact matching	544,990	0.76
English Grade 3 reading	English Grade 4 reading	Single-group design	291,168	0.71
English Grade 4 reading	English Grade 5 reading	Coarsened exact matching	509,216	0.68
English Grade 5 reading	English Grade 6 reading	Coarsened exact matching	559,596	0.75
English Grade 6 reading	English Grade 7 reading	Coarsened exact matching	614,145	0.74
English Grade 7 reading	English Grade 8 reading	Coarsened exact matching	591,859	0.75
Spanish Grade 3 reading	Spanish Grade 4 reading	Single-group design	20,229	0.70
Spanish Grade 4 reading	Spanish Grade 5 reading	Coarsened exact matching	25,903	0.73
Spanish Grade 5 reading	English Grade 6 reading	Coarsened exact matching	25,219	0.69
English Grade 4 writing	Grade 7 writing	Coarsened exact matching	498,391	0.62
Spanish Grade 4 writing	Grade 7 writing	Coarsened exact matching	16,447	0.39
Grade 5 Science	Grade 8 Science	Coarsened exact matching	562,112	0.68

*Correlations are statistical measures of the relationships between scores on separate STAAR assessments. Correlations can range from -1 to 1; high positive values indicate strong positive relationships. For example, students with high STAAR grade 8 mathematics scores tend to have high STAAR Algebra I scores.

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 committee's recommended performance cut scores.

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. Further examination indicated that the nature of the Spanish population in terms of size and changes across grade levels may have resulted in inadequate results. 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. 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.

STAAR 3–8 External Validity Studies

Studies empirically linked student performance on the STAAR grade 8 assessments and the grade 7 writing assessment with external measures linked to postsecondary readiness. EXPLORE and ReadiStep serve as early indicators of postsecondary readiness through links to the ACT and SAT, respectively. Both assessments are typically administered to students in grade 8. Empirically linking STAAR to these assessments helped inform the reasonable ranges for the Level III performance standards and align the performance standards from middle school to high school.

The data for the STAAR grade 8 assessments and the grade 7 writing assessment were based on motivated high-stakes administrations. Performance on EXPLORE and ReadiStep are not tied to academic consequences, potentially resulting in lower student motivation. The representativeness of the assessments in terms of demographics and student proficiency was similar to the student populations. Each linking study had large sample sizes and was based on a single group of students that took both the TAKS tests and the external measure in 2010 or 2011. The STAAR-to-TAKS comparison study provided the results on the STAAR scale. Table A4.3 shows the sample sizes available for each analysis and the correlations between the scores from the linked tests.

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. EXPLORE and ReadiStep were linked to the reference point evaluated for ACT and SAT that represented a 75% chance of earning a C or better in a corresponding college course. Points along the STAAR scales indicating that students would be at least 50% likely to meet or exceed the reference points for EXPLORE and ReadiStep were estimated. Table A4.4 provides the external tests to which the STAAR 3–8 assessments were linked, along with the cut scores examined for those linked tests.

Table A4.3: 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

*Correlations are statistical measures of the relationships between scores on separate STAAR assessments. Correlations can range from -1 to 1; high positive values indicate strong positive relationships. For example, students with high STAAR grade 8 mathematics scores tend to have high STAAR Algebra I scores.

Table A4.4: 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.

The results of the external linking studies were evaluated based on student performance on the STAAR assessments in spring 2012. Specifically, the percent of students at or above the point on the STAAR assessment where there was a 50% chance of being at or above the reference point on the external assessment was computed. Since the majority of students usually attain a grade of C or better in English college courses (Mattern, Patterson, & Kobrin, 2012), the reference points on the ReadiStep Critical Reading and Writing assessments were low (2.5 and 1.0, respectively), resulting in a large percentage of students at or above the reference point. In addition, there were differences in the overlap in content between STAAR assessments and EXPLORE/ReadiStep assessments. These resulted in linking results for ReadiStep Critical Reading and Writing that were non-informative in the neighborhood development process for grade 8 reading and grade 7 writing. However, attaining a grade of C or better in mathematics college courses is more challenging (Mattern, Patterson, & Kobrin, 2012). As a result, the linking study for grade 8 mathematics to ReadiStep mathematics was more informative in the development of the neighborhood.

STAAR Vertical Scale Studies

The STAAR 3–8 vertical scale study was used to establish a common scale across grades 3–8 mathematics and reading, as well as across grades 3–5 Spanish reading. By having the assessments on a common scale across grade levels, comparisons can be made in terms of student performance from year to year. The vertical scale study determined the difference in test difficulty between adjacent grades for STAAR 3–8 reading and mathematics assessments. The relationship between grades on the vertical scale helped inform the alignment of performance standards.

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 test forms).

Item-response theory, specifically the Rasch model, was used to place all items and student proficiencies on the same scale within a content area. Table A4.5 shows the cumulative vertical scale linking constants.

Table A4.5: Cumulative Vertical Scale Constants for STAAR 3–8

Grades	STAAR 3–8 Mathematics	STAAR 3–8 Reading	STAAR 3–5 Spanish Reading
8	0	0	-
7	-0.4388	-0.2101	-
6	-0.9130	-0.6679	-
5	-1.6506	-1.0586	0
4	-2.1030	-1.3854	-0.2700
3	-2.7895	-2.0057	-0.8513

The development of the neighborhoods for grades 3–8 was informed by the vertical scale. The neighborhoods for grade 8 reading and mathematics, which were based on links to the EOC final performance standards, were mapped down from grade 7 to grade 3 using the vertical scale constants.

The vertical scale technical report will be available on the TEA website in spring 2013 at http://www.tea.state.tx.us/student.assessment/reports/.

STAAR 3–8 to TAKS Comparisons

The comparison between STAAR 3–8 assessments and TAKS assessments (bridge studies) was based on an average of two methodologies: an empirical method (item response theory) and impact data. For most of the STAAR-to-TAKS comparison studies in grades 3–8, STAAR items were embedded in the spring 2011 TAKS administration. The Rasch model resulted in the

STAAR and TAKS items and proficiency levels being placed on the same Rasch scale. The TAKS Met Standard proficiency level, θ_{ms}, was identified by finding the STAAR proficiency level that equaled the TAKS Met Standard proficiency level.

For the grades 4 and 7 writing assessments, a separate STAAR stand-alone field test was administered rather than embedding STAAR items in the TAKS writing assessments. Therefore, the data-collection design for the writing assessments was based on a single-group design. Students who took TAKS writing assessments in spring 2011 and the STAAR stand-alone field test in spring 2011 were included in the sample. An equipercentile linking was conducted to identify the TAKS Met Standard proficiency level, θ_{ms}, on the STAAR assessment.

The empirical result was evaluated with respect to trends in TAKS impact data and the impact data for the 2012 STAAR assessments. This resulted in an average between the empirical method and the impact data method. An example of the result of the bridge study is provided in Table A4.6. The values in Table A4.6 do not represent study results; they are provided for illustrative purposes.

Table A4.6: Example STAAR-to-TAKS Comparison

STAAR Grade 8 Reading Proficiency Estimates	TAKS Grade 8 Performance Standard
…	…
-0.5	
-0.4	
-0.3	TAKS Met Standard
-0.2	
-0.1	
…	…

Additional information regarding the bridge studies can be found on the TEA website at www.tea.state.tx.us/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=2147504930&libID=2147504925.

STAAR 3–8 Relation to NAEP Impact Data

Results from the National Assessment of Educational Progress (NAEP) were considered during the creation of neighborhoods that were used by the educator committees during the STAAR 3–8 standard-setting process. Comparisons between STAAR performance and NAEP performance were used to evaluate the rigor of the STAAR performance standards relative to standards established for national assessment instruments.

When creating STAAR 3–8 neighborhoods, STAAR 3–8 performance and NAEP performance were compared within a content area. Specifically, the NAEP assessments administered at grade 8 were compared to STAAR grade 8 assessments in the same content areas for reading, mathematics, and science. NAEP grade 8 writing was compared to STAAR grade 7 writing. NAEP

grade 8 geography was compared to STAAR grade 8 social studies. The NAEP assessments administered at grade 4 were compared to STAAR grade 4 assessments in the same content areas for reading, mathematics, and writing. Texas-specific and national data were used for all assessments except for NAEP grade 8 geography, where only national data were available. A single group of common examinees was not available for these analyses, so NAEP results were given limited weight during the STAAR 3–8 standard-setting process. Figures A4.1–A4.8 show the NAEP impact data that was considered during the development of the neighborhoods for standard-setting.

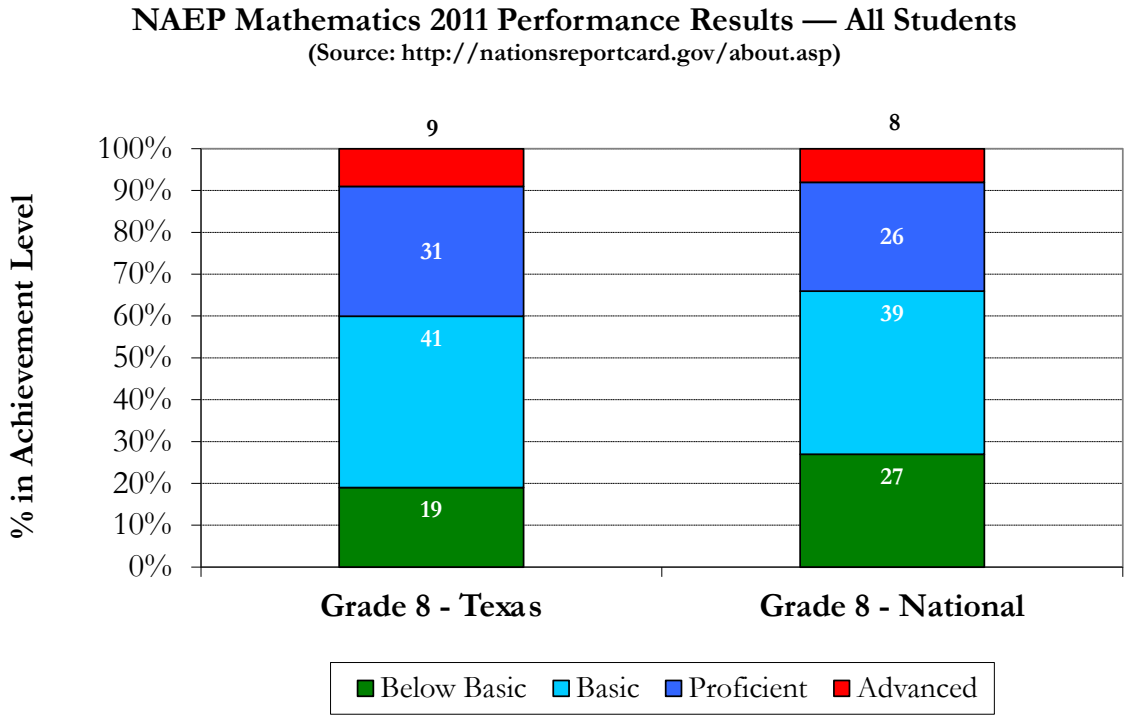


Figure A4.1 – NAEP Grade 8 Mathematics Impact Data

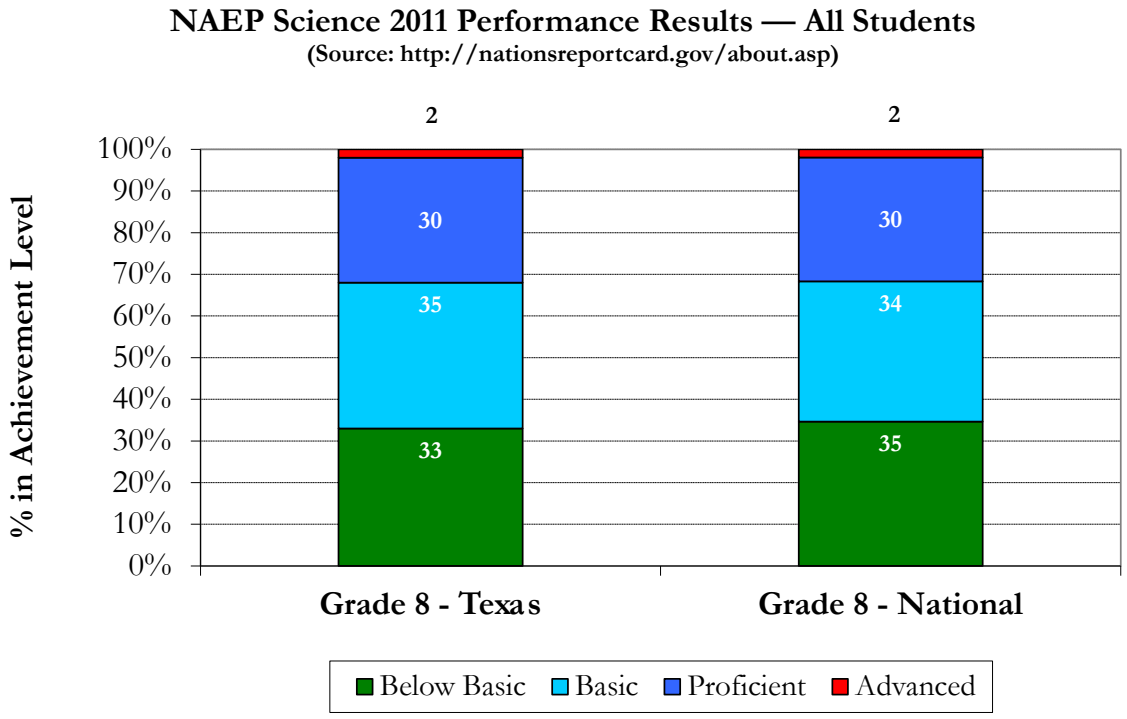


Figure A4.2 – NAEP Grade 8 Science Impact Data

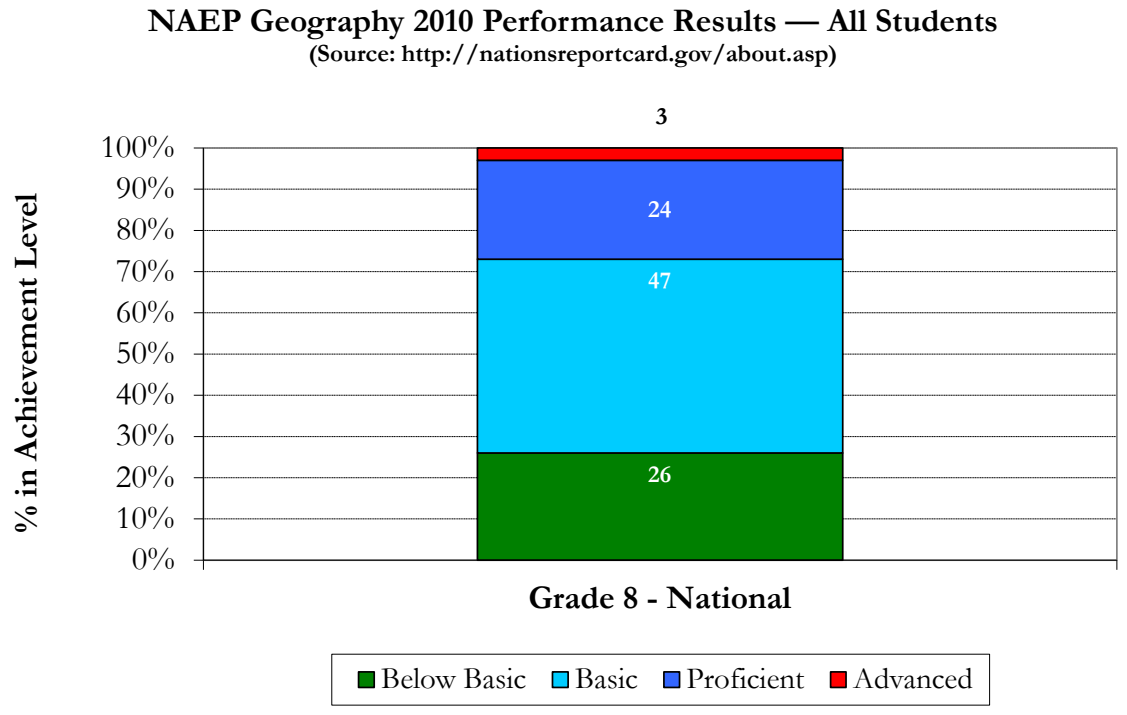


Figure A4.3 – NAEP Grade 8 Geography Impact Data

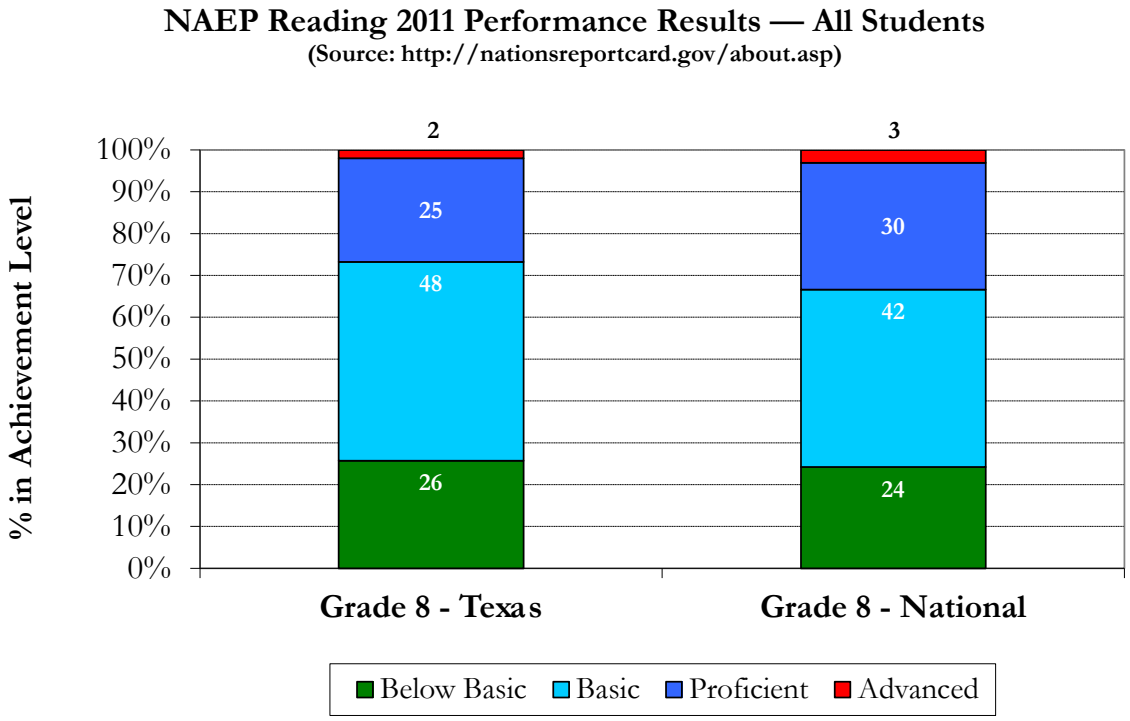


Figure A4.4 – NAEP Grade 8 Reading Impact Data

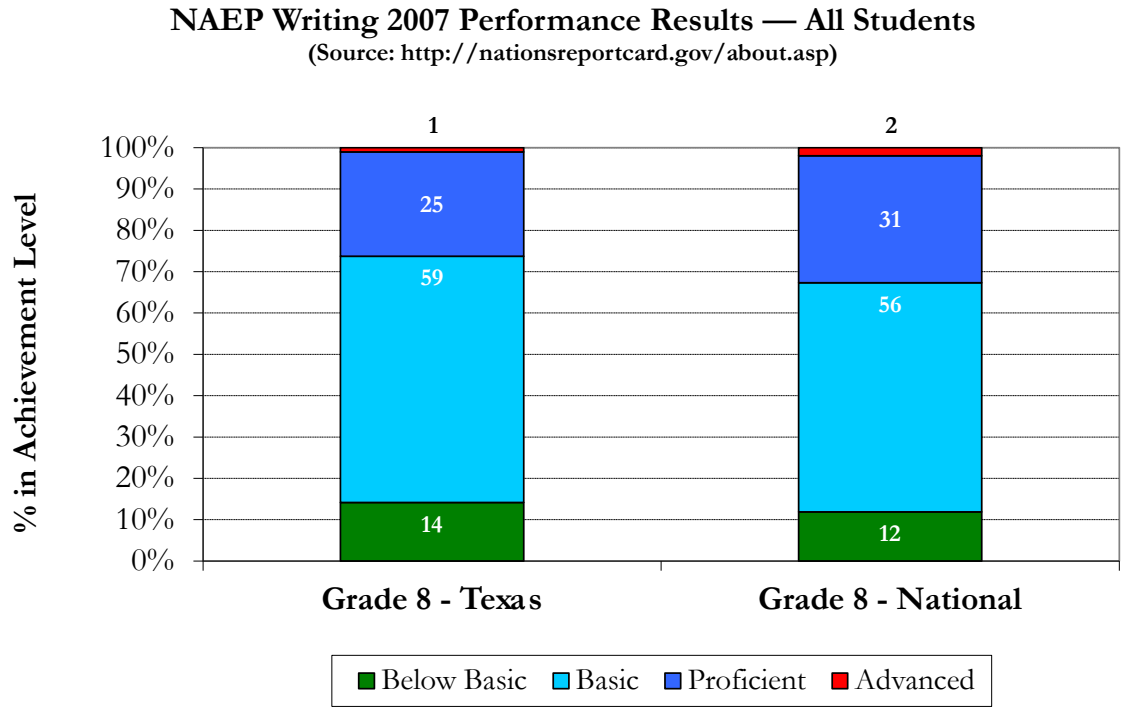


Figure A4.5 – NAEP Grade 8 Writing Impact Data

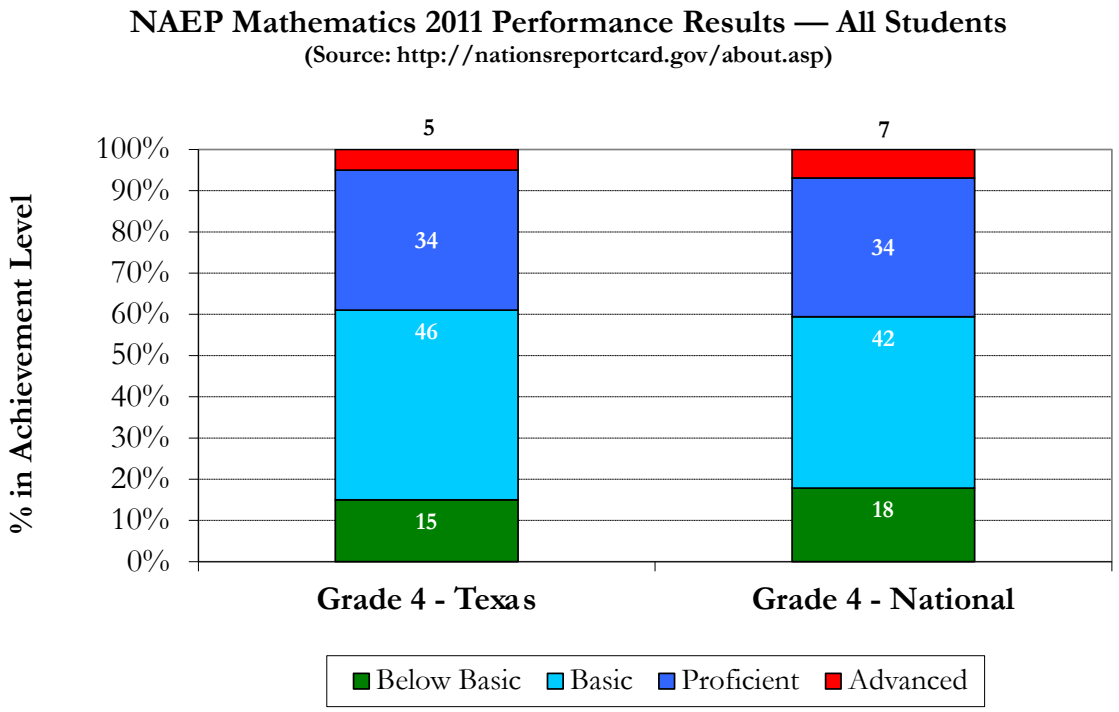


Figure A4.6 – NAEP Grade 4 Mathematics Impact Data

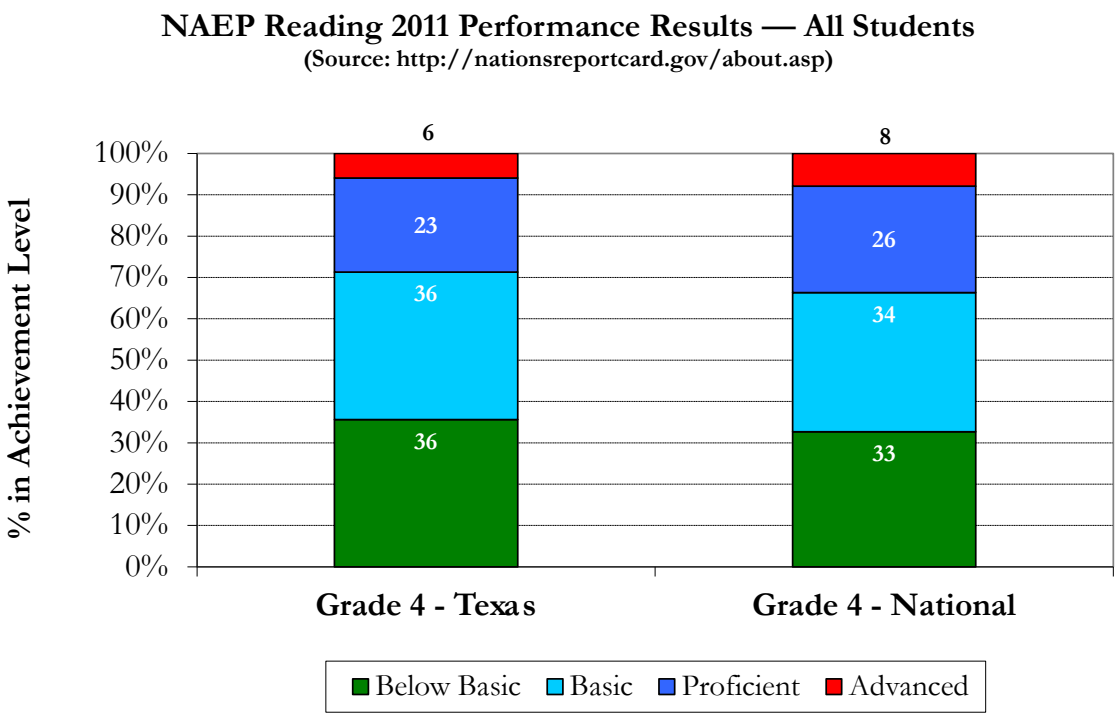


Figure A4.7 – NAEP Grade 4 Reading Impact Data

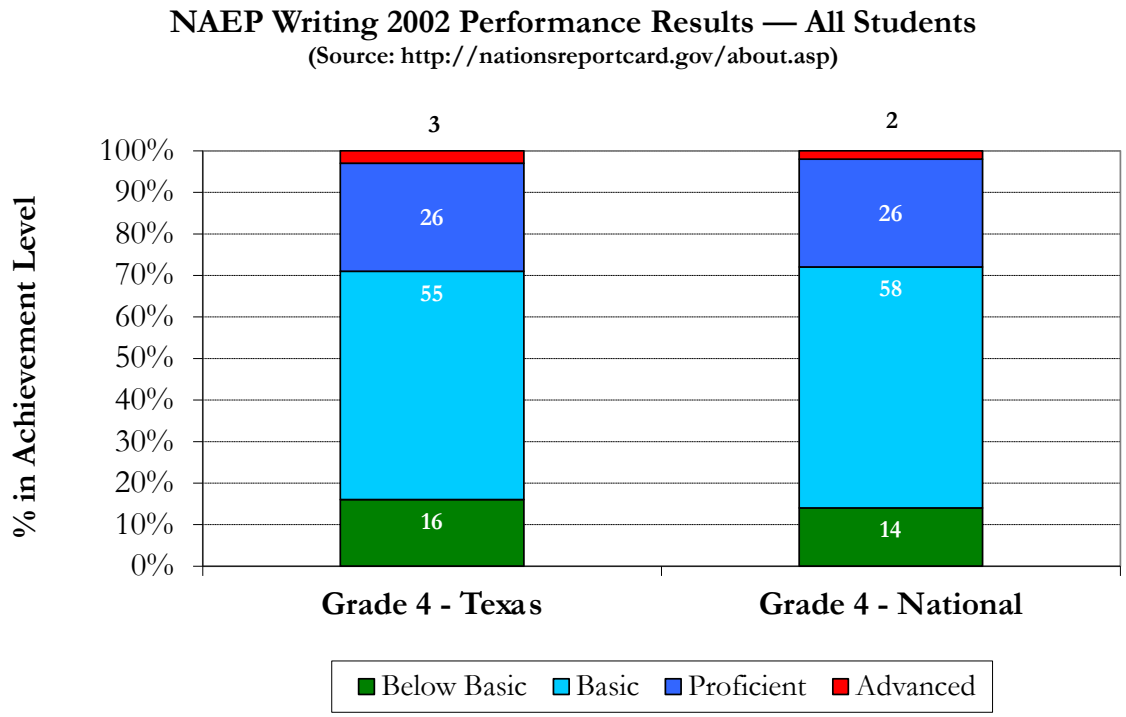


Figure A4.8 – NAEP Grade 4 Writing Impact Data

Appendix 5: PLD Meeting Process Evaluation Summary

At the close of the PLD meetings, educators were asked to provide confidential input on their experience with PLD development. The evaluation covered the training tasks, the comfort level of the participants in expressing their opinions, the time allotted for the meeting, the degree to which the committee could distinguish among the PLDs at each performance level, and the likelihood that committee members would participate again or recommend participation to their colleagues.

Because the evaluations were distributed at the close of the cross-content articulation of the PLDs, they reflect the aggregated opinions of all committees for a particular content area: mathematics, English, science, and social studies.

A summary of the responses from the STAAR EOC PLD meetings is listed in the chart below.

Question	Content Area	Number of Respondents	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree	Omit
1. During training my role on the committee was made clear	Math	19	89%	11%				
	English	12	100%					
	Science	24	92%	8%				
	Soc.Studies	18	83%	17%				
2. The conference leaders clearly explained the tasks I needed to complete	Math	19	89%	11%				
	English	12	100%					
	Science	24	96%	4%				
	Soc.Studies	18	89%	11%				
3. I felt comfortable expressing my opinions.	Math	19	89%	11%				
	English	12	100%					
	Science	24	100%					
	Soc.Studies	18	89%	11%				
4. Everyone was given the opportunity to express his/her opinion.	Math	19	100%					
	English	12	100%					
	Science	24	100%					
	Soc.Studies	18	94%	6%				
5. I could clearly distinguish between levels of achievement represented in the performance level descriptors.	Math	19	74%	26%				
	English	12	100%					
	Science	24	79%	17%				4%
	Soc.Studies	18	89%	11%				
6. There was sufficient time to complete the assigned tasks.	Math	19	95%			5%		
	English	12	67%	25%		8%		
	Science	24	96%	4%				
	Soc.Studies	18	89%	11%				
7. I would be willing to assist again on a similar task.	Math	19	100%					
	English	12	100%					
	Science	24	100%					
	Soc.Studies	18	100%					
8. I would recommend this activity to a colleague.	Math	19	100%					
	English	12	100%					
	Science	24	92%	4%			4%	
	Soc.Studies	18	100%					

A summary of the responses from the STAAR 3–8 PLD meetings is listed in the chart below.

Question	Content Area	Number of Respondents	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree	Omit
1. During training my role on the committee was made clear	Math	37	84%	16%				
	Reading	39	92%	8%				
	Writing	14	93%	7%				
	Science	11	91%	9%				
	Soc.Studies	8	100%					
2. The conference leaders clearly explained the tasks I needed to complete	Math	37	81%	19%				
	Reading	39	97%	3%				
	Writing	14	100%					
	Science	11	91%	9%				
	Soc.Studies	8	100%					
3. I felt comfortable expressing my opinions.	Math	37	86%	14%				
	Reading	39	87%	10%	3%			
	Writing	14	93%	7%				
	Science	11	73%	27%				
	Soc.Studies	8	100%					
4. Everyone was given the opportunity to express his/her opinion.	Math	37	92%	8%				
	Reading	39	95%	5%				
	Writing	14	100%					
	Science	11	100%					
	Soc.Studies	8	100%					
5. I could clearly distinguish between levels of achievement represented in the performance level descriptors.	Math	37	70%	30%				
	Reading	39	92%	8%				
	Writing	14	86%	14%				
	Science	11	82%	18%				
	Soc.Studies	8	100%					
6. There was sufficient time to complete the assigned tasks.	Math	37	89%	11%				
	Reading	39	100%					
	Writing	14	100%					
	Science	11	100%					
	Soc.Studies	8	100%					
7. I would be willing to assist again on a similar task.	Math	37	95%	5%				
	Reading	39	97%	3%				
	Writing	14	100%					
	Science	11	91%	9%				
	Soc.Studies	8	100%					
8. I would recommend this activity to a colleague.	Math	37	92%	8%				
	Reading	39	97%	3%				
	Writing	14	100%					
	Science	11	91%	9%				
	Soc.Studies	8	100%					

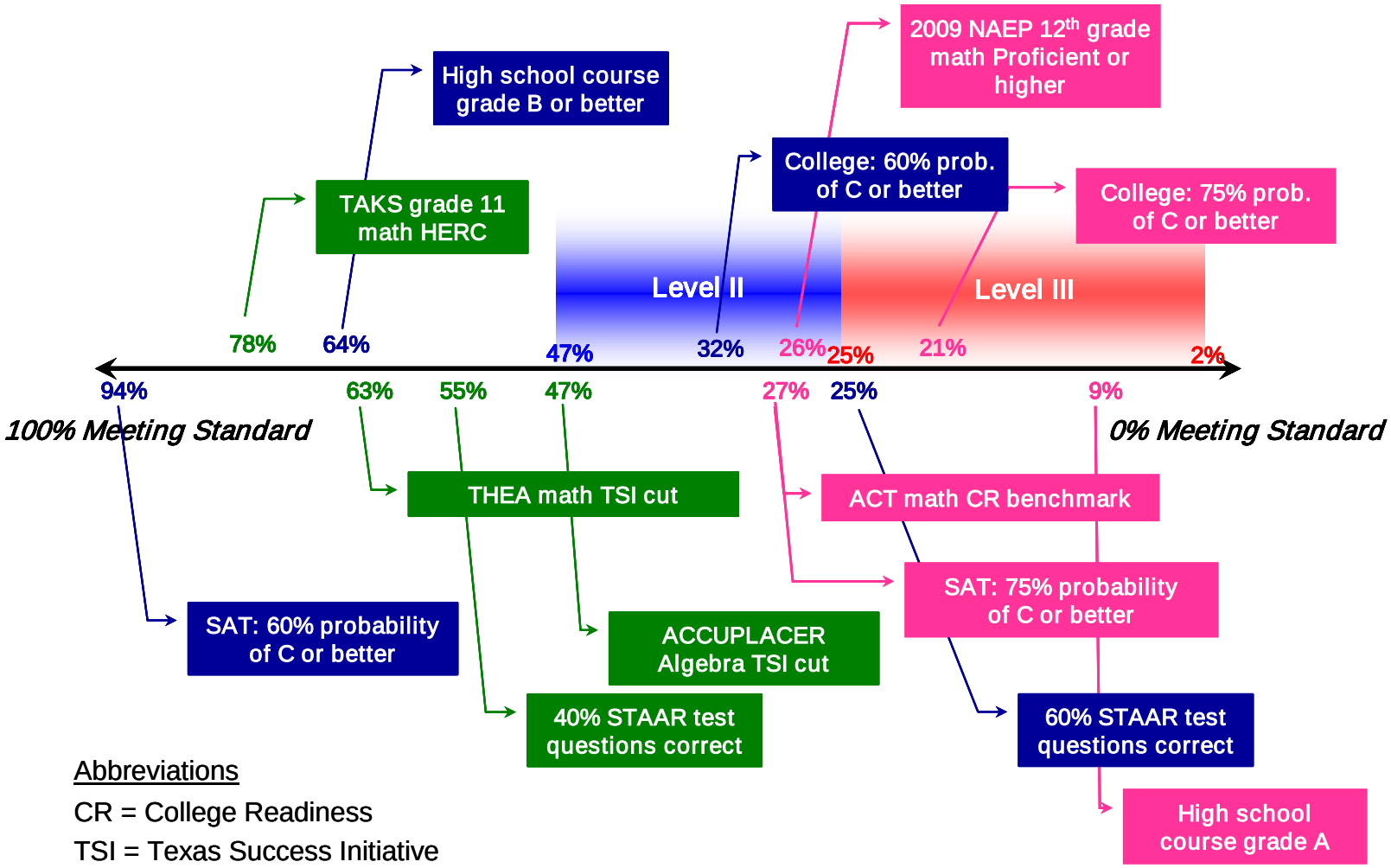
Appendix 6: Policy Committee Members

A complete listing of all policy committee members, including their names, positions, and affiliations at the time of the policy committee meeting, is given in the table below.

Name	Position and Affiliation
Dana Bedden	Superintendent, Irving ISD
Michael Bettersworth	Associate Vice Chancellor, Technology Advancement, Texas State Technical College
Reece Blincoe	Superintendent, Brownwood ISD
Bobby Blount	Board of Trustees, Northside Independent School District
Courtney Boswell	Senior Policy Analyst, Senate Education Committee
Von Byer	Committee Director, Senate Education Committee
Jesus Chavez	Superintendent, Round Rock ISD
Patti Clapp	Managing Director, Patti Clapp Consulting
David Dunn	Executive Director, Texas Charter Schools Association
Andrew Erben	President, Texas Institute for Education Reform
Kalese Hammonds	Public Education Advisor, Office of the Governor
Don Hernandez	Principal, Garland ISD
Troy Johnson	Associate Vice President of Academic Affairs, University of North Texas
Sandy Kress	Senior Counsel, Akin, Grump, Strauss, Hauer, and Feld
Caasi Lamb	Policy Analyst, Office of the Lt. Governor
Russell Lowery-Hart	Vice President of Academic Affairs, Amarillo College
Robert Nelsen	President, University of Texas-Pan American
Donna Newman	Executive Director of Middle School Campus Administration, North East ISD
Anne Poplin	Executive Director, Education Service Center, Region IX
Todd Rogers	Principal, Northwest ISD
Rod Schroder	Superintendent, Amarillo ISD; Past President and Member, Texas School Alliance
Andrea Sheridan	Senior Policy Advisor, Office of the Speaker of the House
Jennifer Shiess	Manager, Legislative Budget Board
Stephanie Stoebe	Teacher, Round Rock ISD
Jeri Stone	Executive Director/General Counsel, Texas Classroom Teachers Association
Rod Townsend	Superintendent, Denton ISD
Lori Vetters	President, Inland Resources
Jenna Watts	Committee Director, House Public Education Committee

Appendix 7: STAAR EOC Empirical Studies Number Lines

STAAR ALGEBRA II

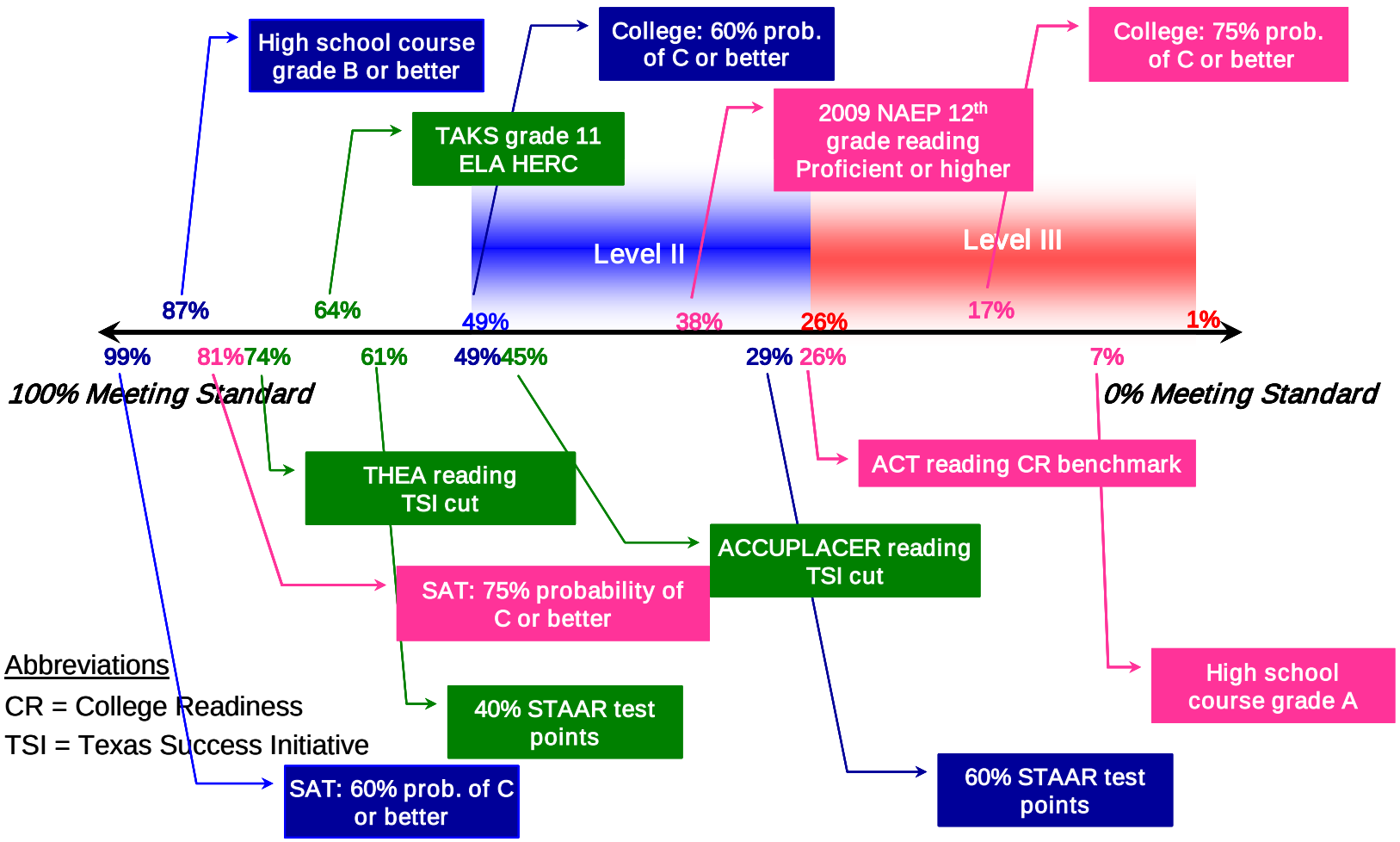


Abbreviations

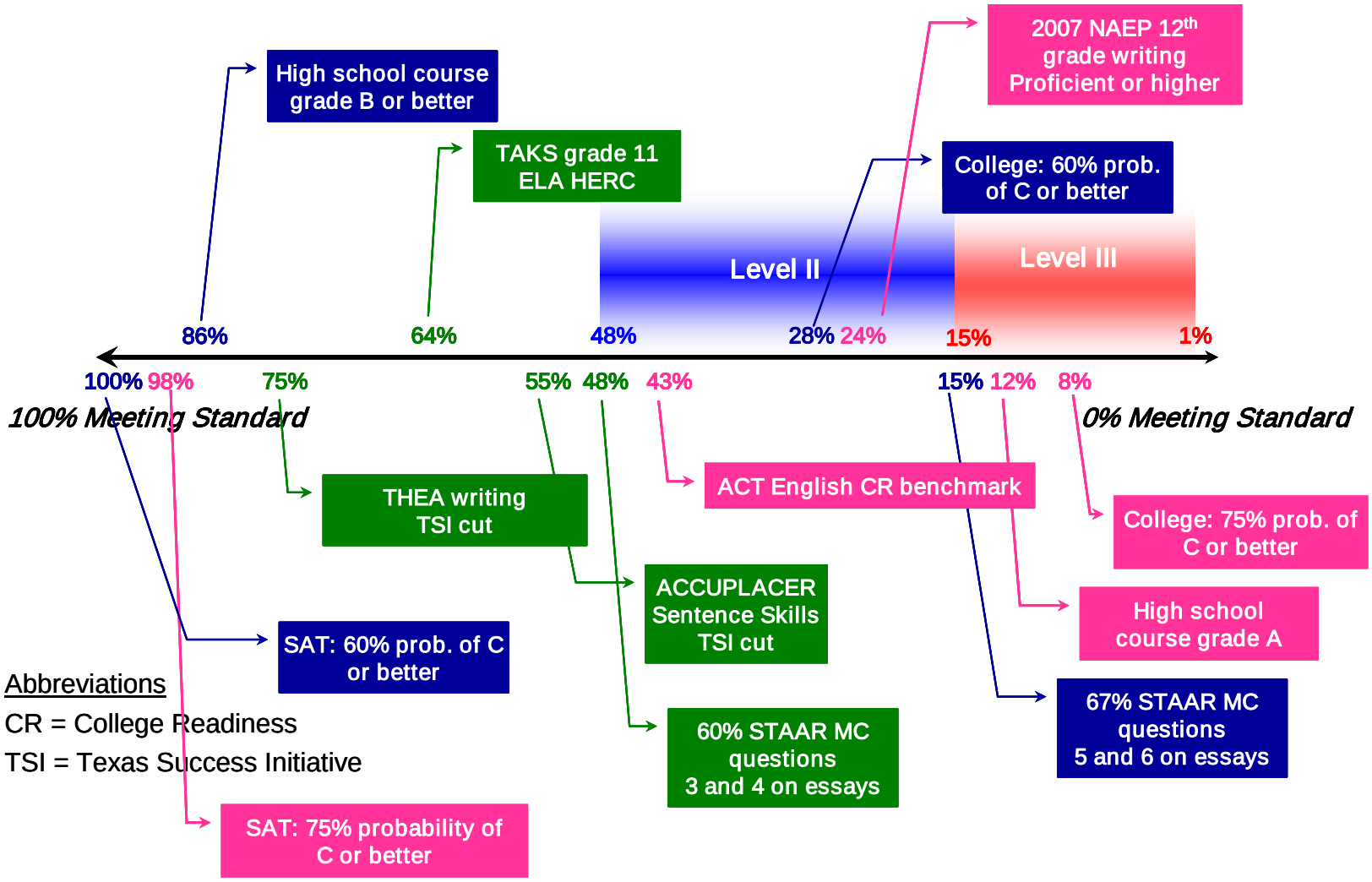
CR = College Readiness

TSI = Texas Success Initiative

STAAR ENGLISH III READING



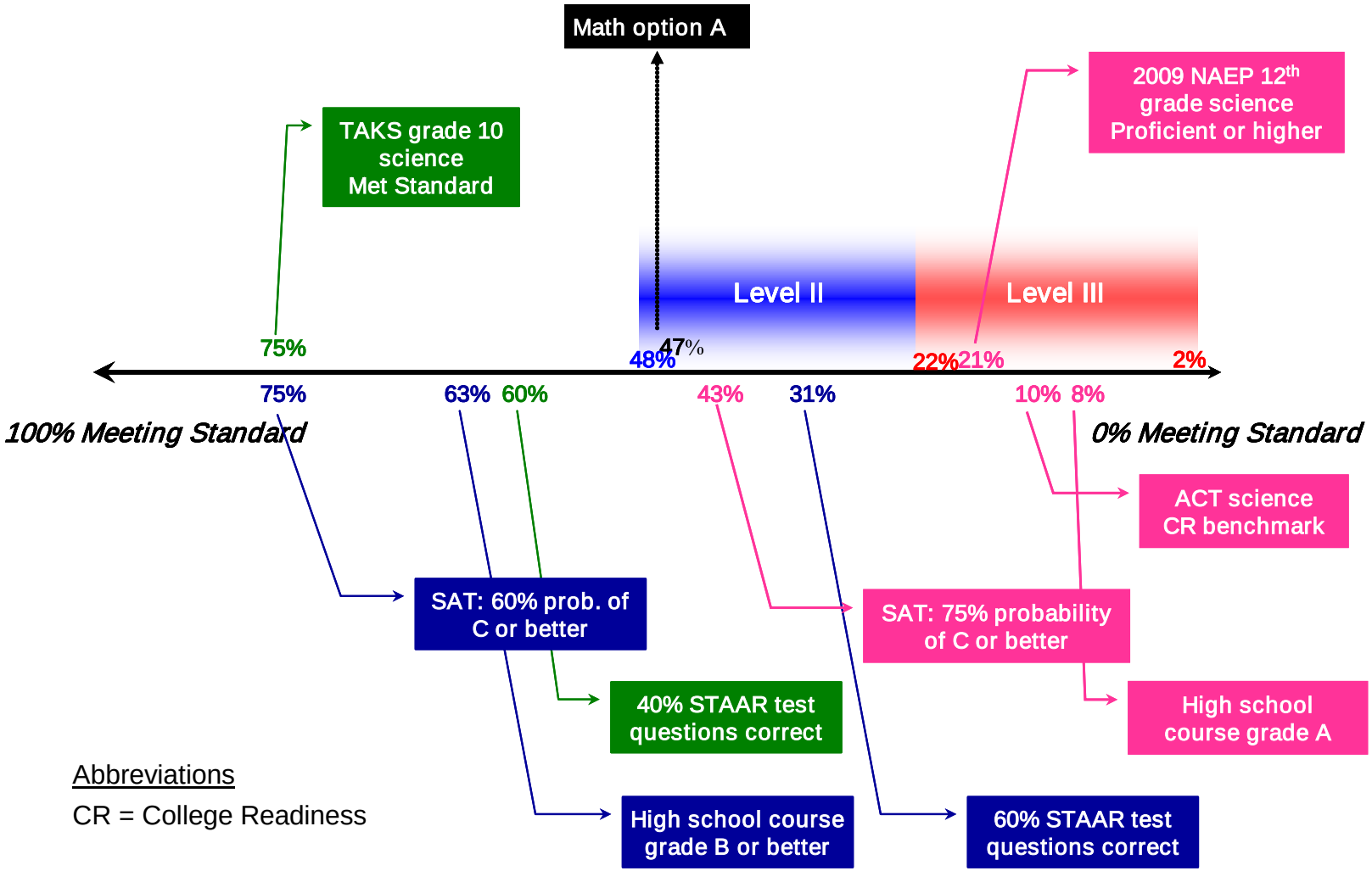
STAAR ENGLISH III WRITING



Abbreviations

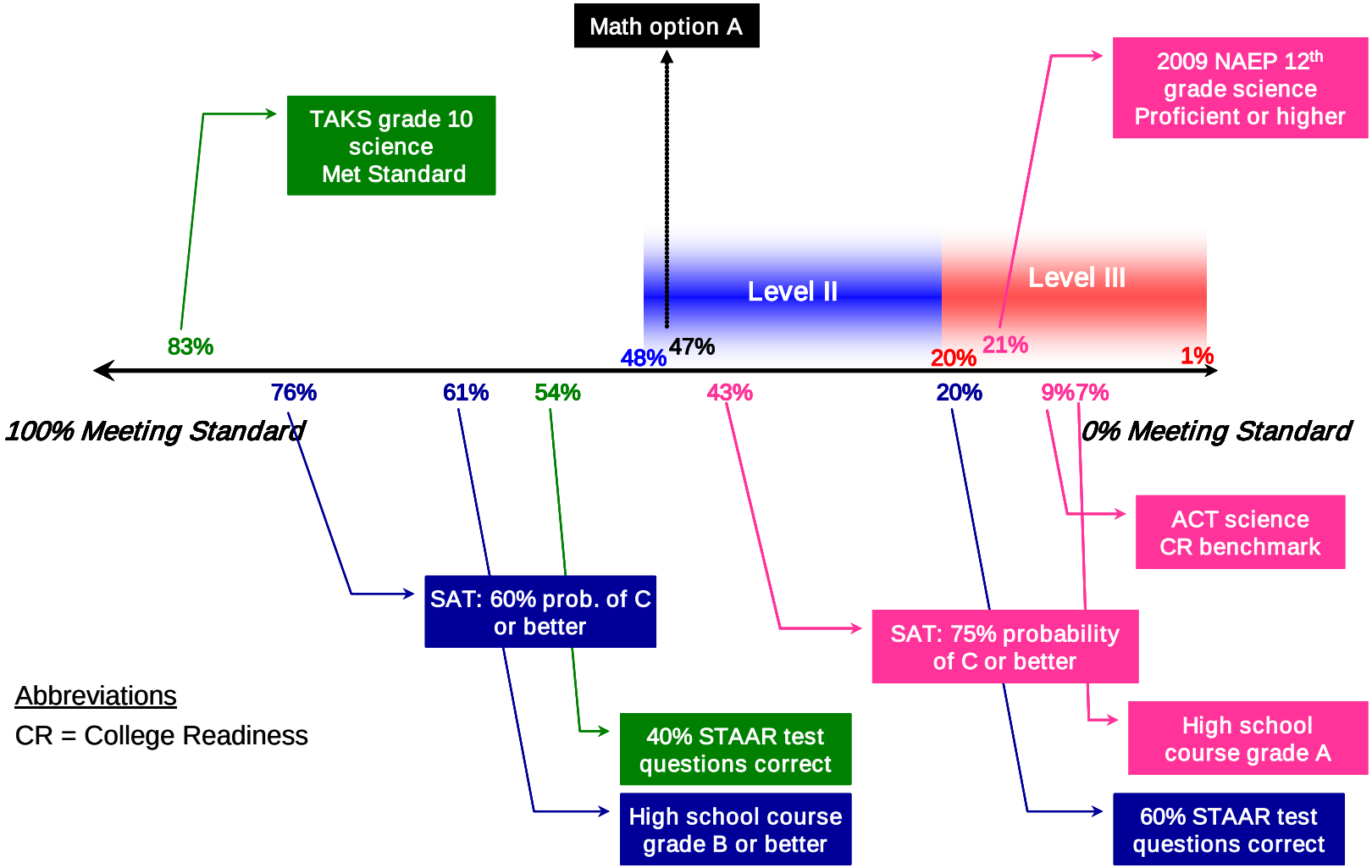
CR = College Readiness

TSI = Texas Success Initiative



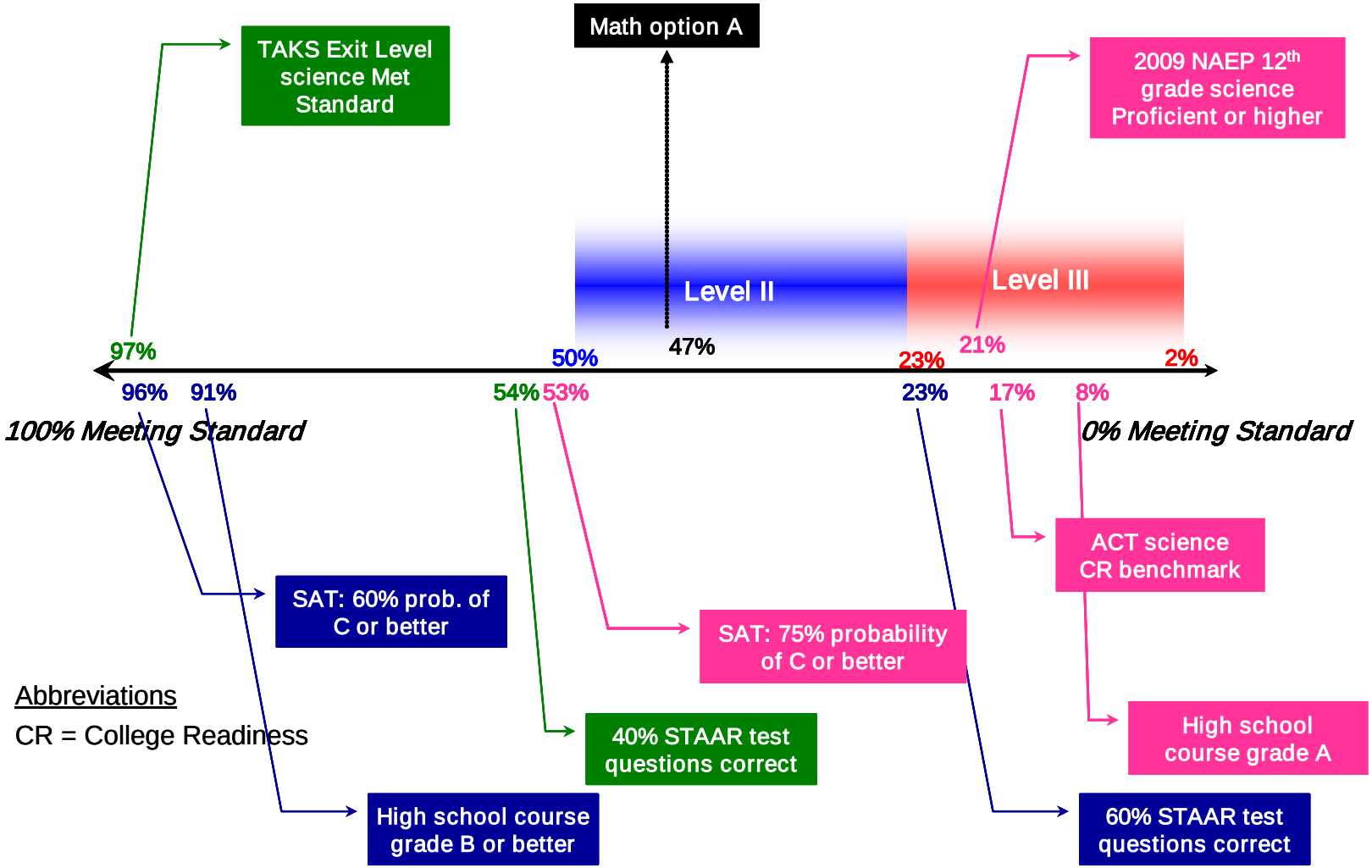
Abbreviations

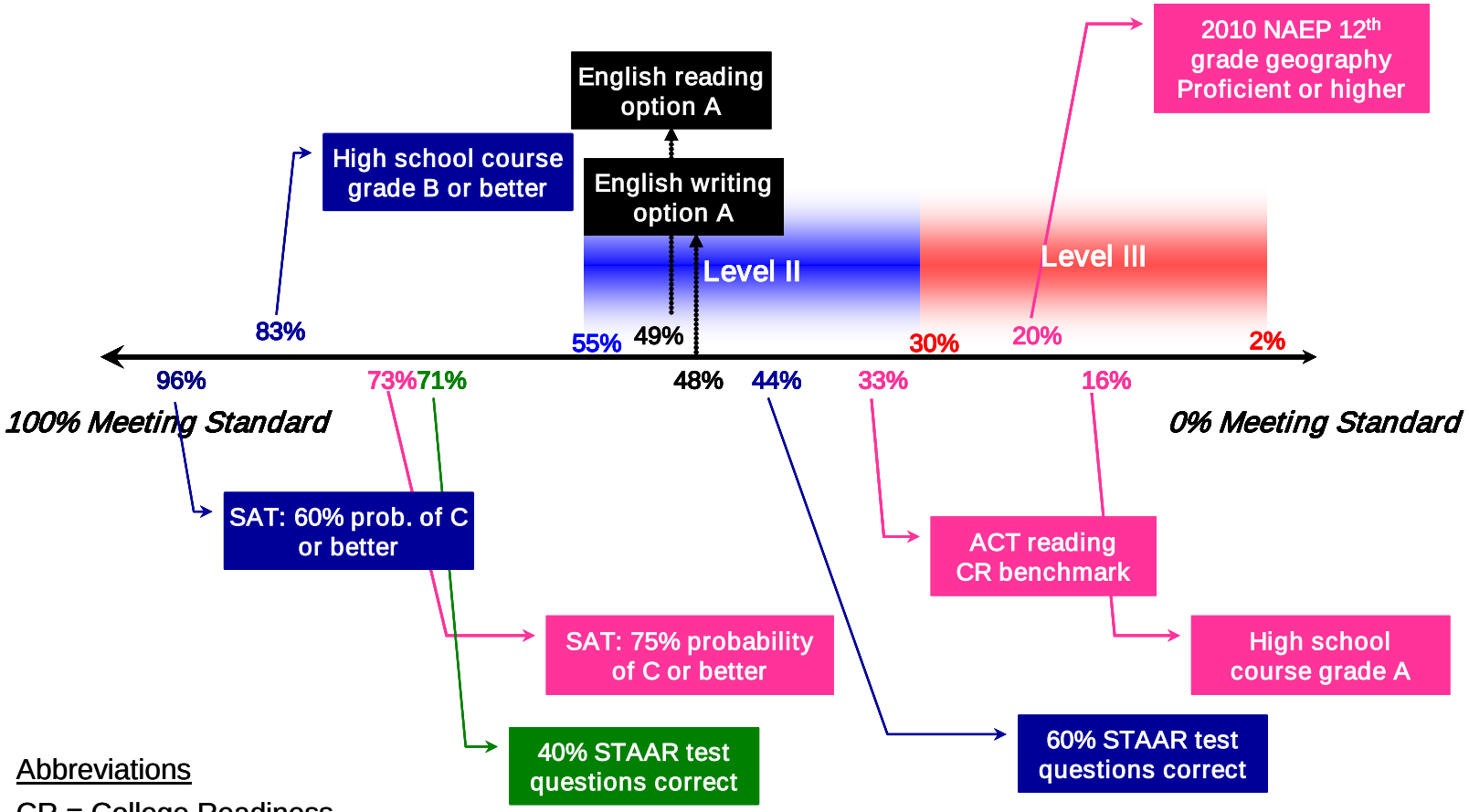
CR = College Readiness



Abbreviations

CR = College Readiness

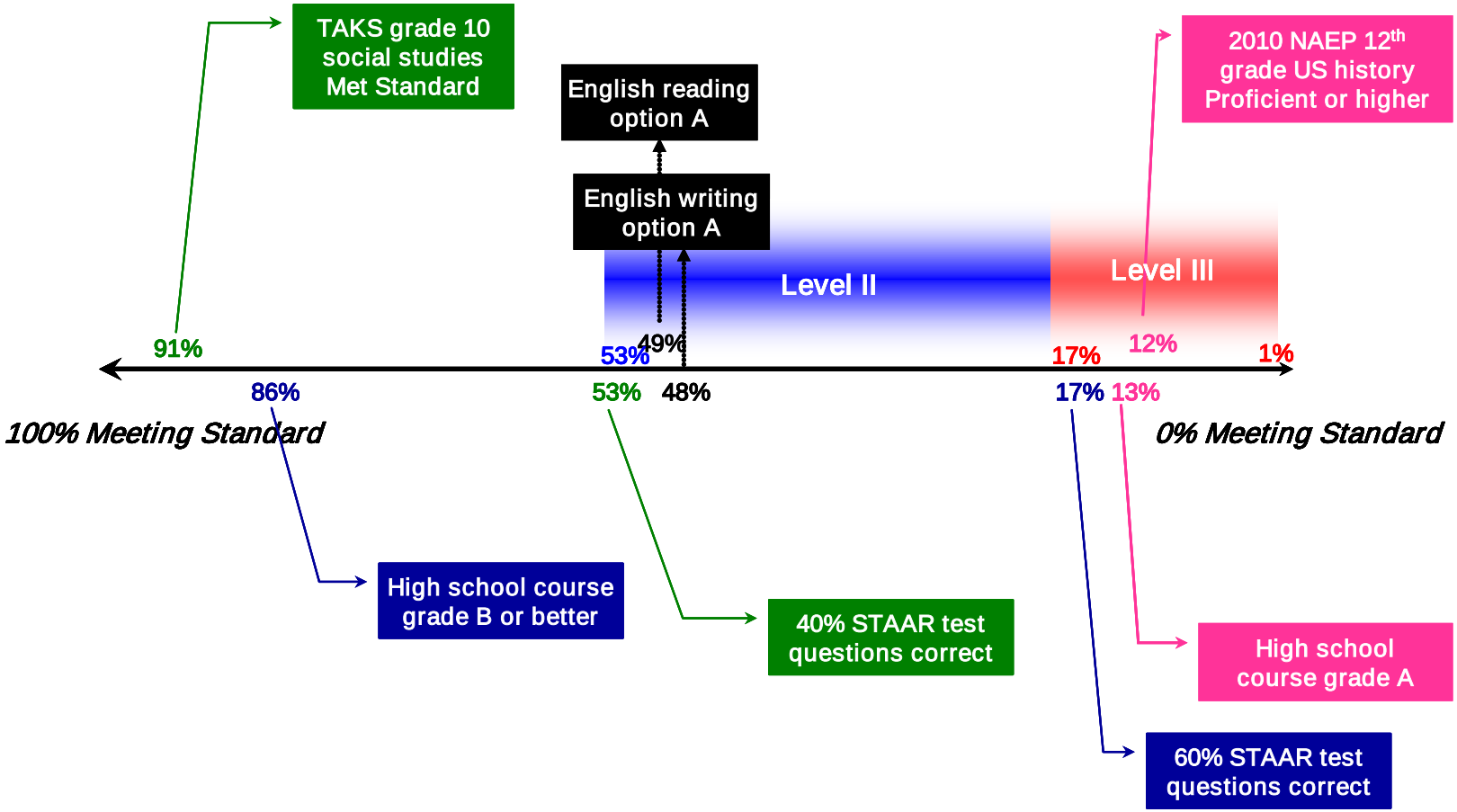

Math option A
TAKS Exit Level science Met Standard
2009 NAEP 12th grade science Proficient or higher
Level II
Level III
97%
47%
50%
21%
23%
2%
96%
91%
54%
53%
23%
17%
8%
100% Meeting Standard
0% Meeting Standard
ACT science CR benchmark
SAT: 60% prob. of C or better
SAT: 75% probability of C or better
High school course grade A
40% STAAR test questions correct
High school course grade B or better
60% STAAR test questions correct
Abbreviations
CR = College Readiness



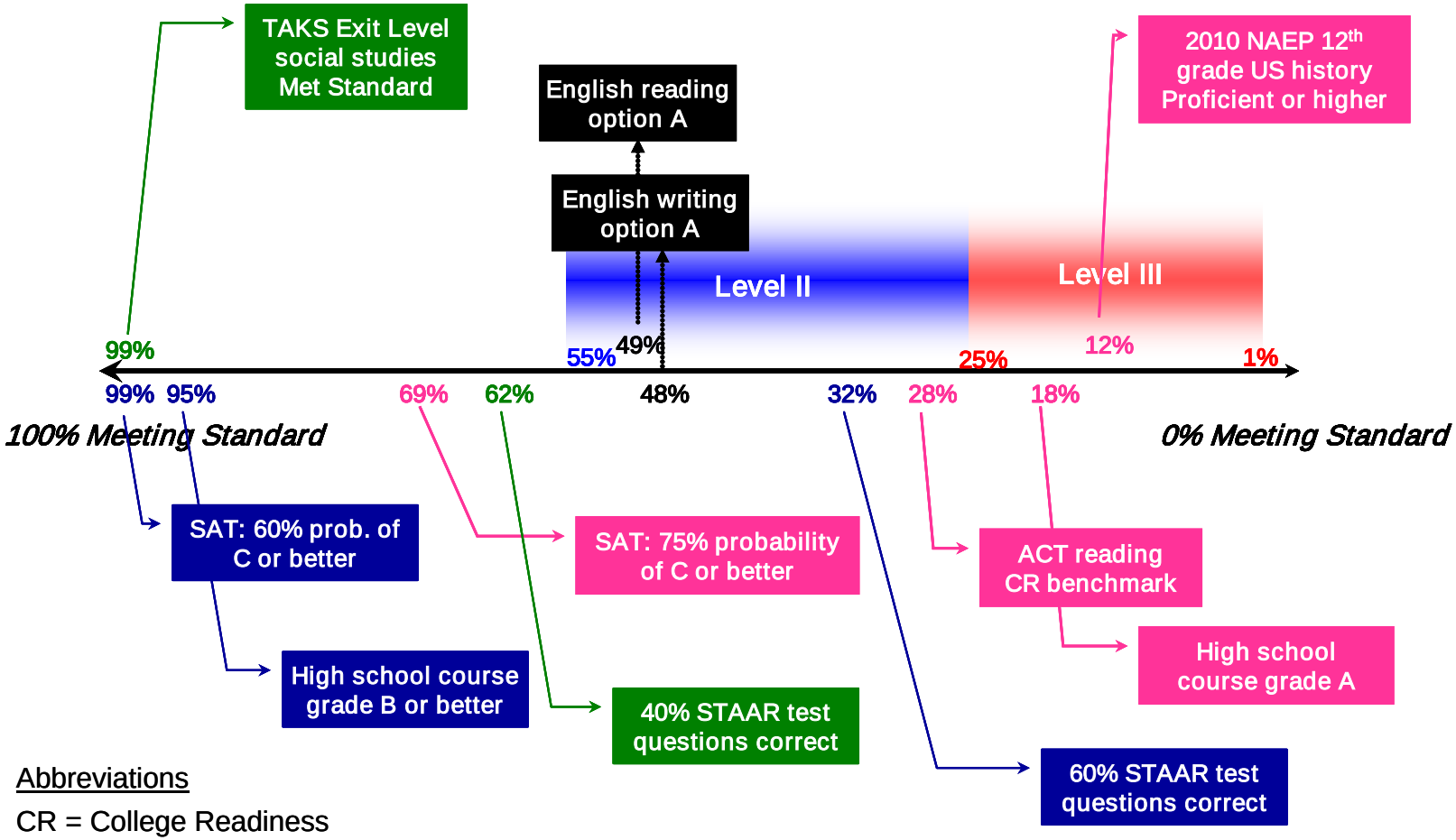
Abbreviations

CR = College Readiness

STAAR WORLD HISTORY



STAAR U.S. HISTORY



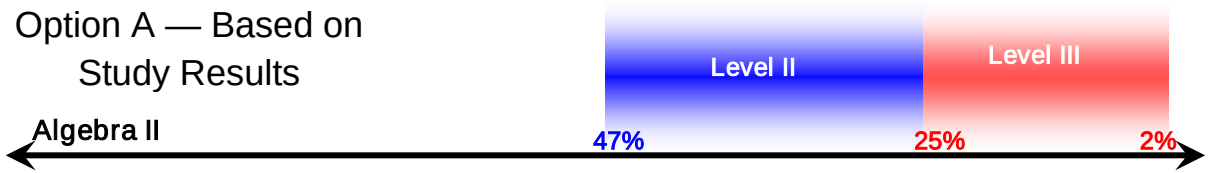
Abbreviations

CR = College Readiness

Appendix 8: STAAR EOC Neighborhood Options

PART 1 – STAAR ALGEBRA II NEIGHBORHOOD OPTIONS

Option A — Based on Study Results



100% Meeting Standard *0% Meeting Standard*

Up to 62% of students may be in Level II or above under high-stakes conditions.

Option B — Adjusted for Motivation

100% Meeting Standard *0% Meeting Standard*

Up to 47% of students may be in Level II or above under high-stakes conditions.

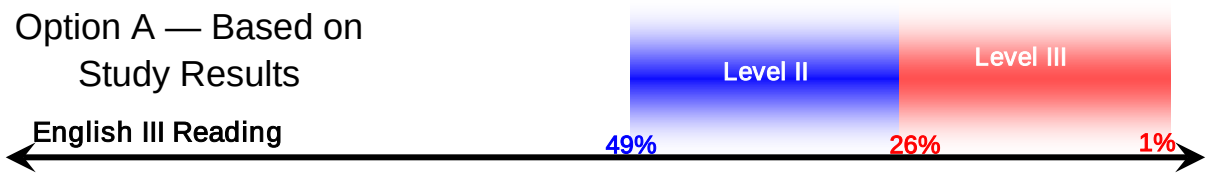
Option C — Additional Adjustment

100% Meeting Standard *0% Meeting Standard*

Up to 45% of students may be in Level II or above under high-stakes conditions.

PART 1 – STAAR ENGLISH III READING NEIGHBORHOOD OPTIONS

Option A — Based on Study Results



100% Meeting Standard *0% Meeting Standard*

Up to 69% of students may be in Level II or above under high-stakes conditions.

Option B — Adjusted for Motivation

100% Meeting Standard *0% Meeting Standard*

Up to 49% of students may be in Level II or above under high-stakes conditions.

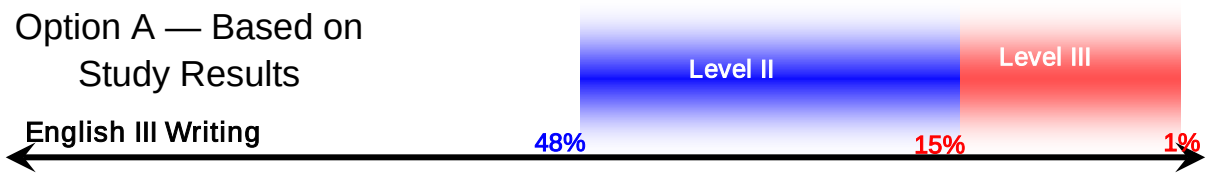
Option C — Additional Adjustment

100% Meeting Standard *0% Meeting Standard*

Up to 45% of students may be in Level II or above under high-stakes conditions.

PART 1 – STAAR ENGLISH III WRITING NEIGHBORHOOD OPTIONS

Option A — Based on Study Results



100% Meeting Standard *0% Meeting Standard*

Up to 68% of students may be in Level II or above under high-stakes conditions.

Option B — Adjusted for Motivation

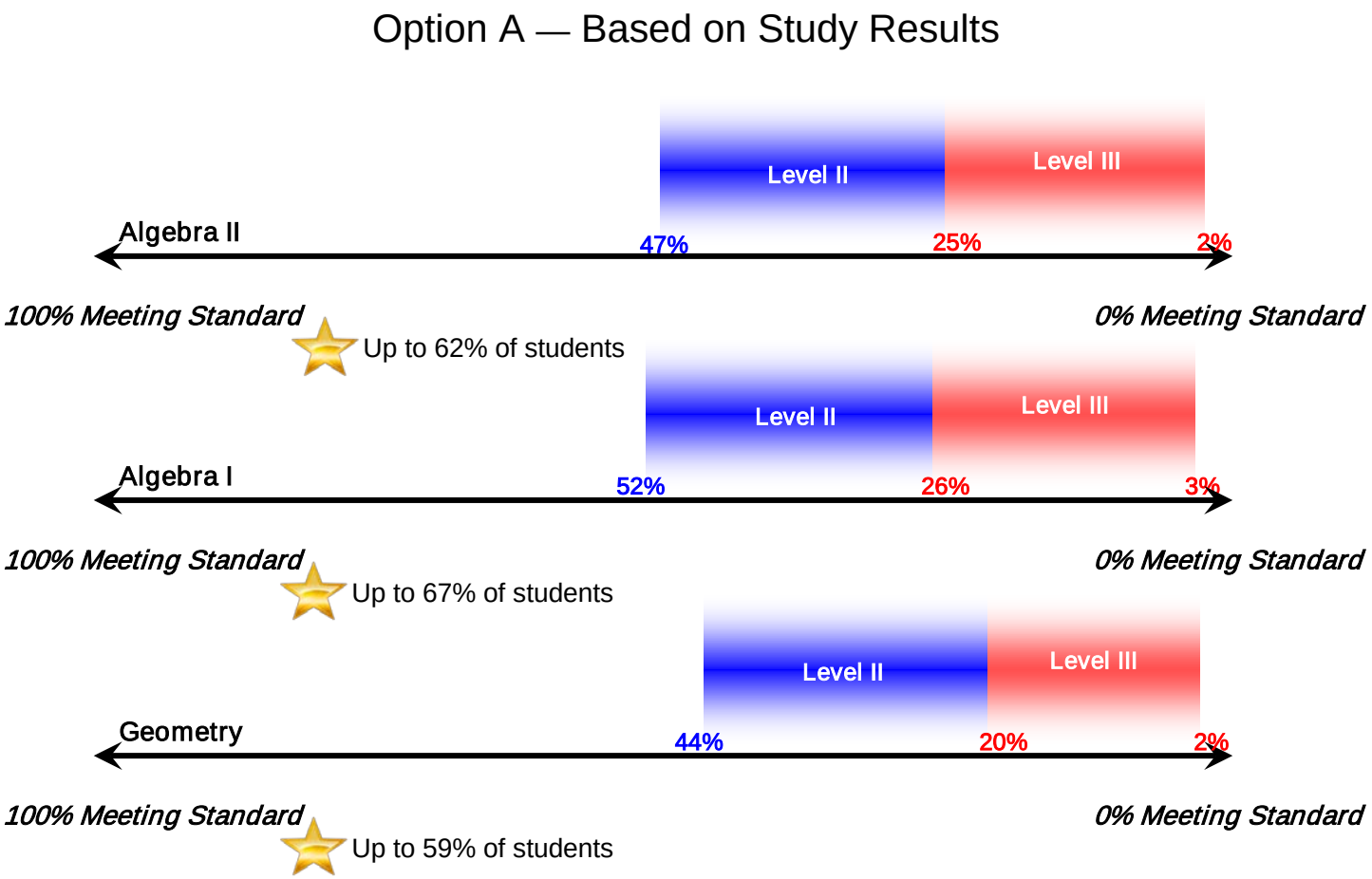
100% Meeting Standard *0% Meeting Standard*

Up to 48% of students may be in Level II or above under high-stakes conditions.

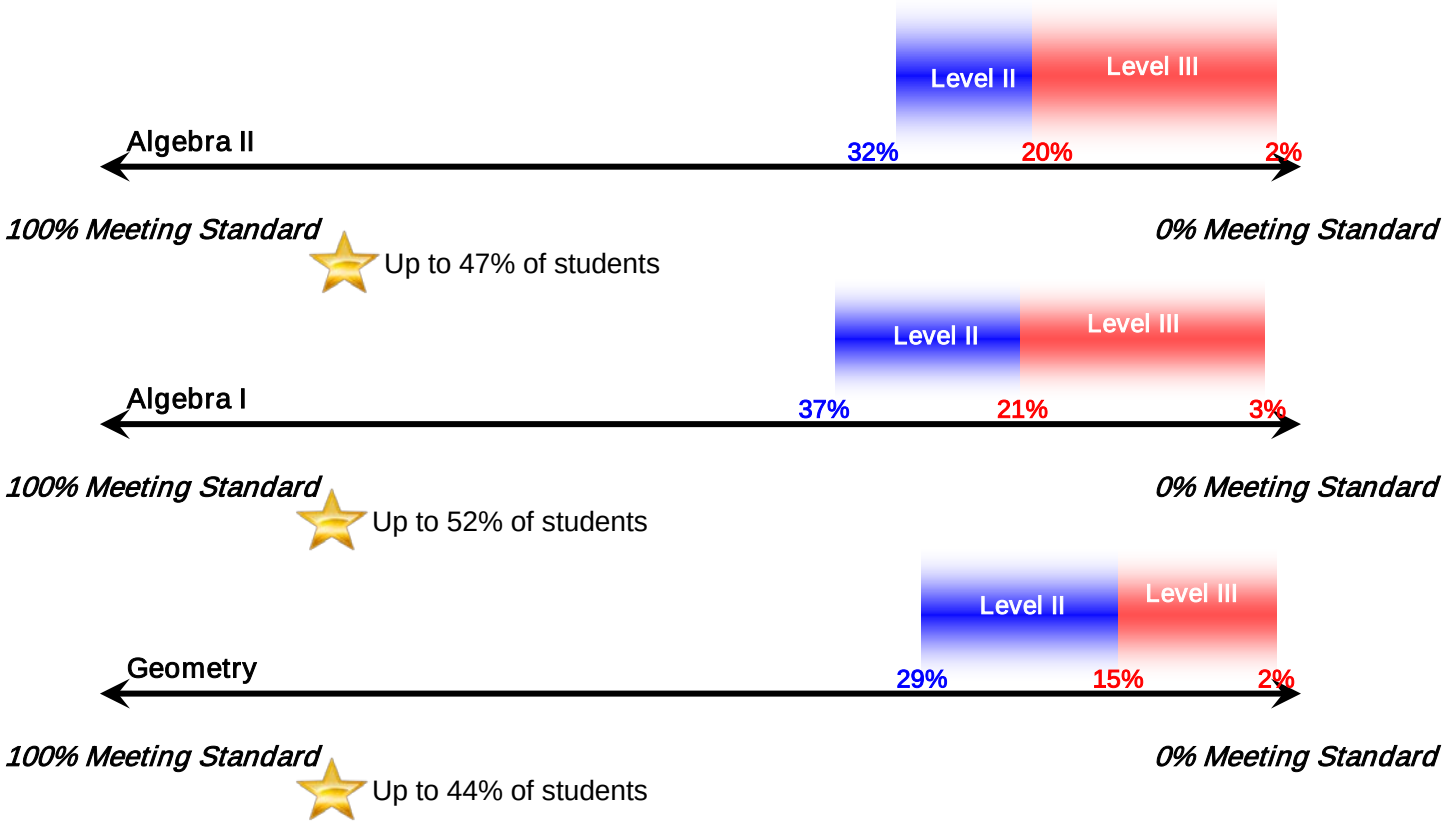
Option C — Additional Adjustment

100% Meeting Standard *0% Meeting Standard*

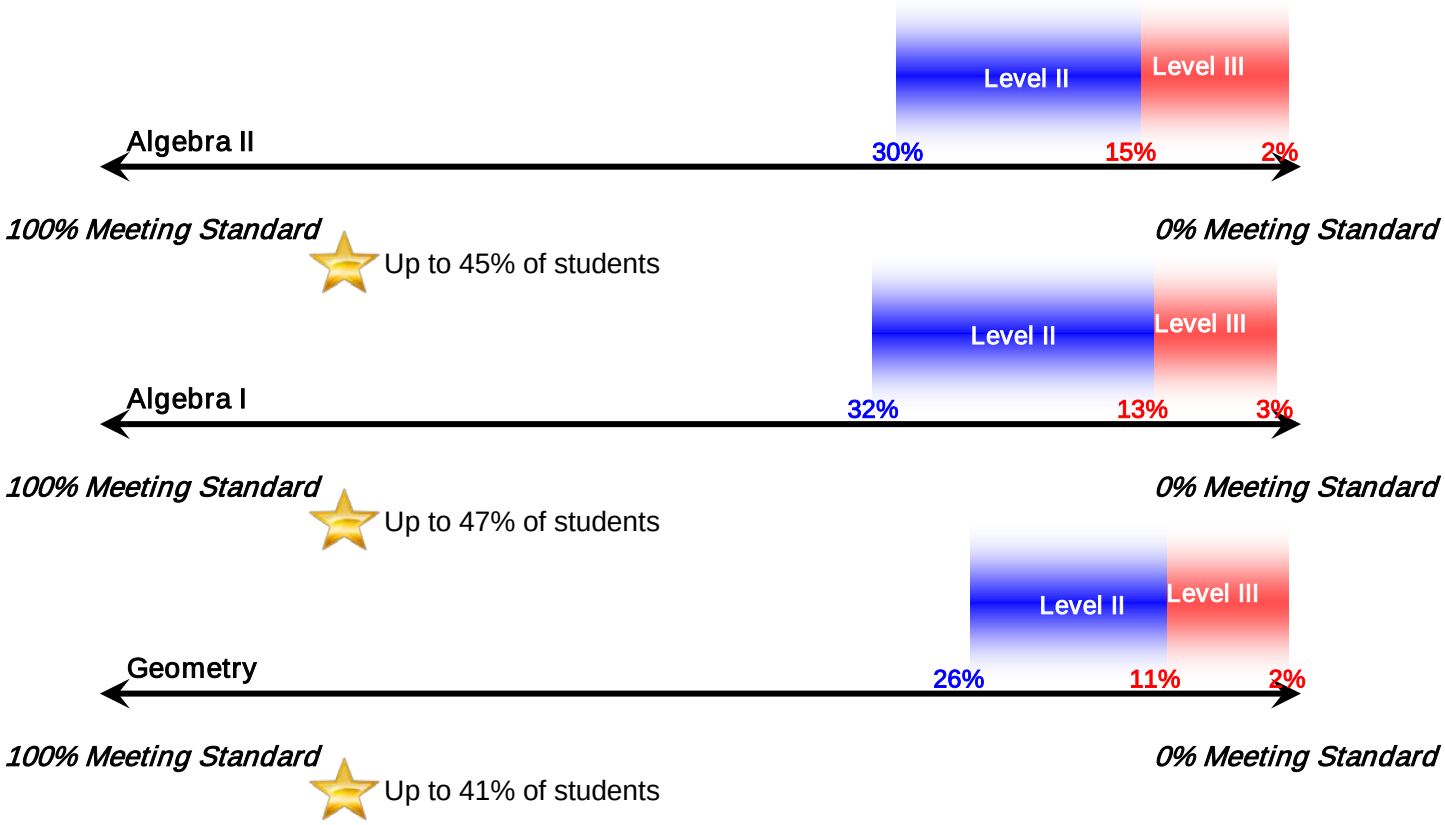
Up to 45% of students may be in Level II or above under high-stakes conditions.

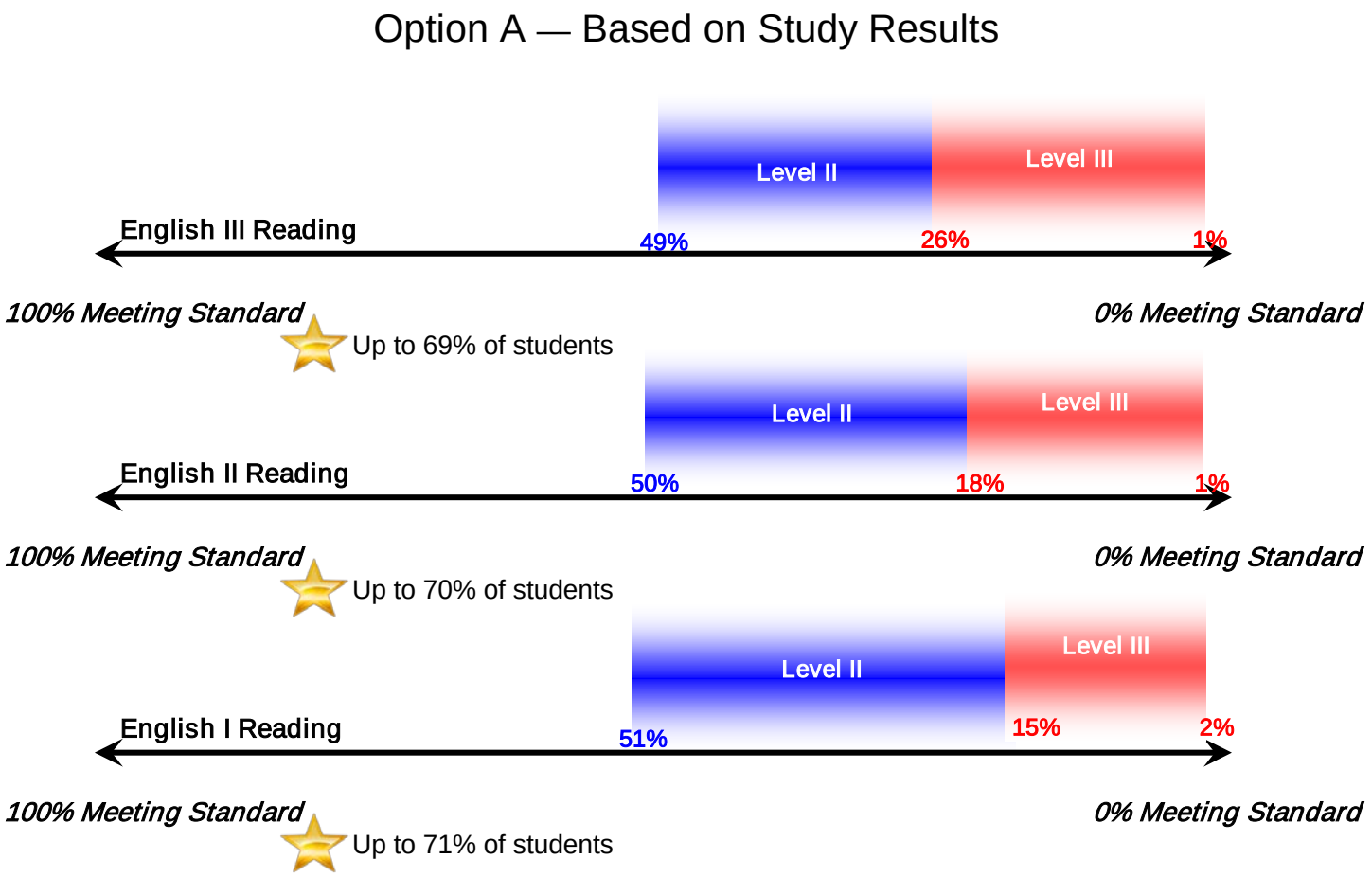

Option A — Based on Study Results
Level II
Level III
Algebra II
47%
25%
2%
100% Meeting Standard
0% Meeting Standard
Up to 62% of students
Level II
Level III
Algebra I
52%
26%
3%
100% Meeting Standard
0% Meeting Standard
Up to 67% of students
Level II
Level III
Geometry
44%
20%
2%
100% Meeting Standard
0% Meeting Standard
Up to 59% of students

Option B — Adjusted for Motivation

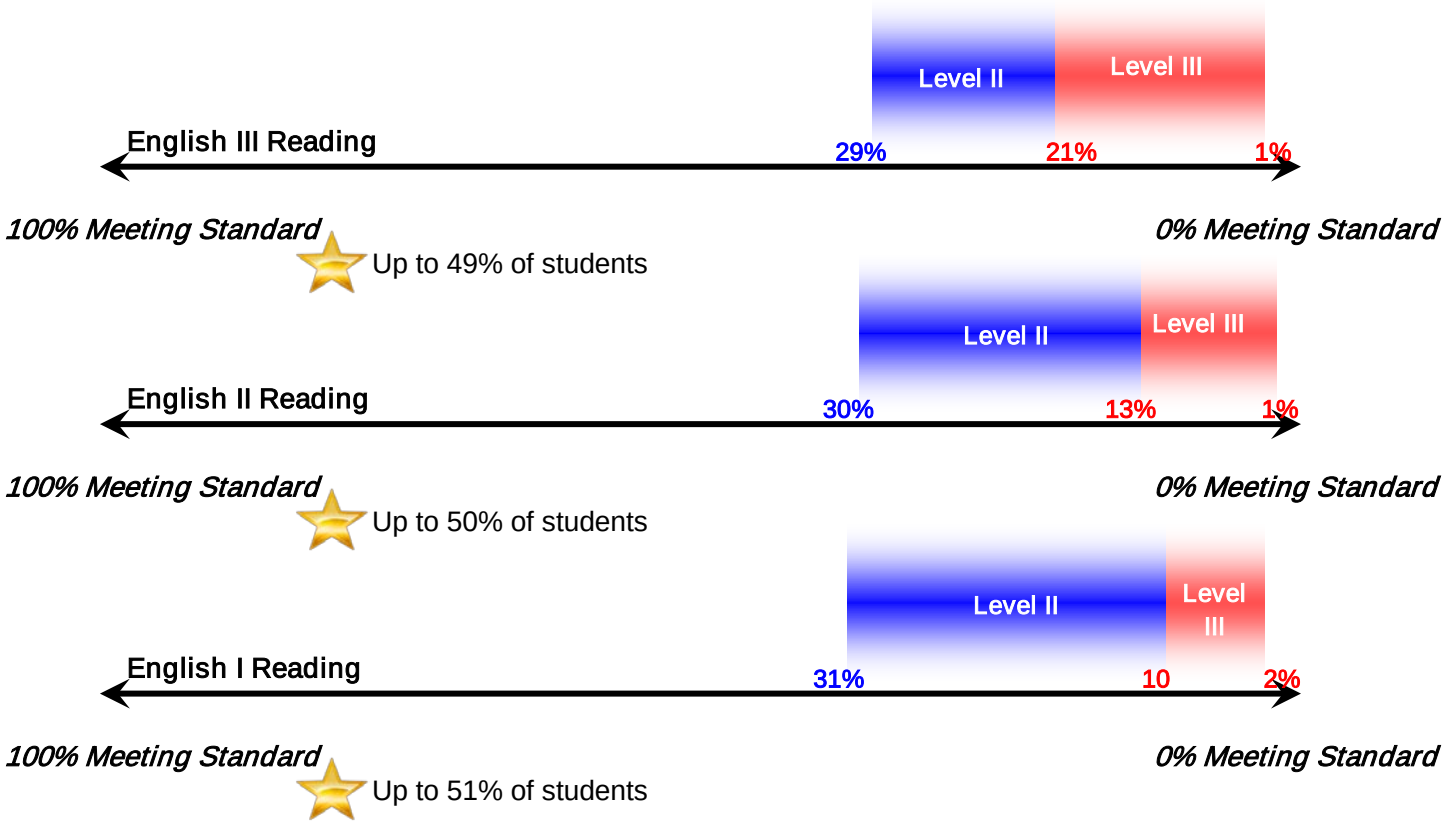


Option C — Additional Adjustment

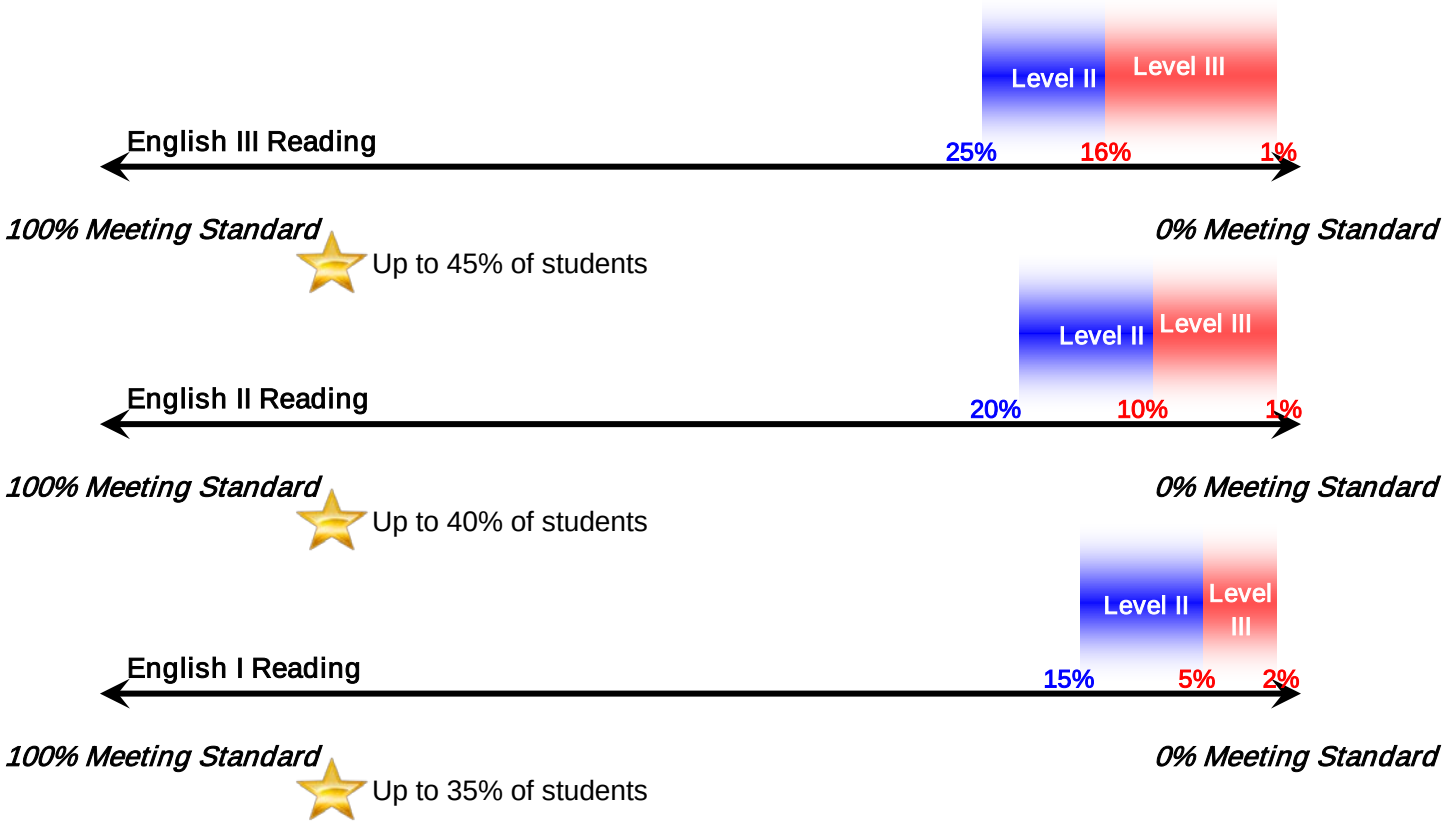



Option A — Based on Study Results
Level II
Level III
English III Reading
49%
26%
1%
100% Meeting Standard
0% Meeting Standard
Up to 69% of students
Level II
Level III
English II Reading
50%
18%
1%
100% Meeting Standard
0% Meeting Standard
Up to 70% of students
Level II
Level III
English I Reading
51%
15%
2%
100% Meeting Standard
0% Meeting Standard
Up to 71% of students

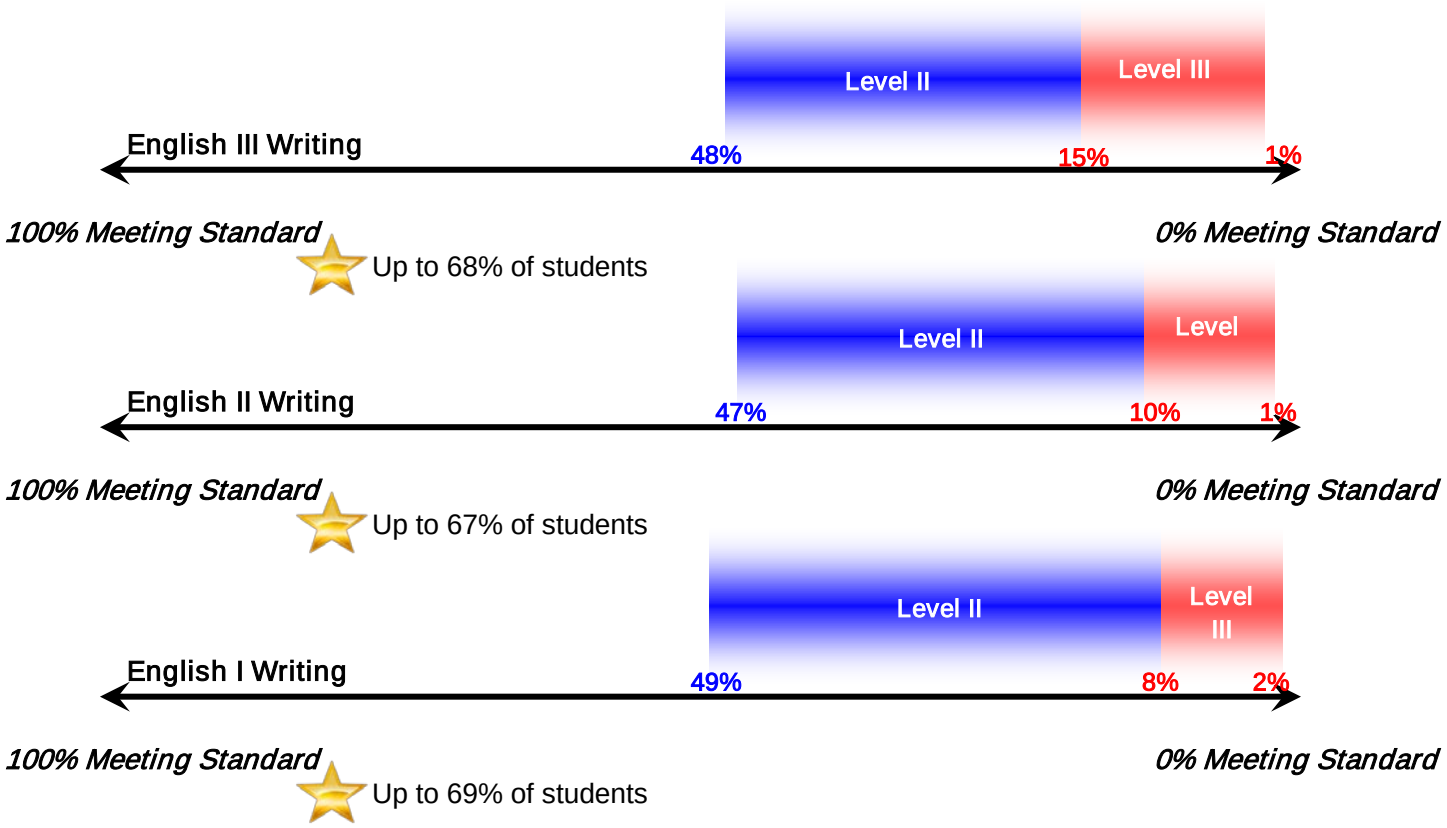
Option B — Adjusted for Motivation



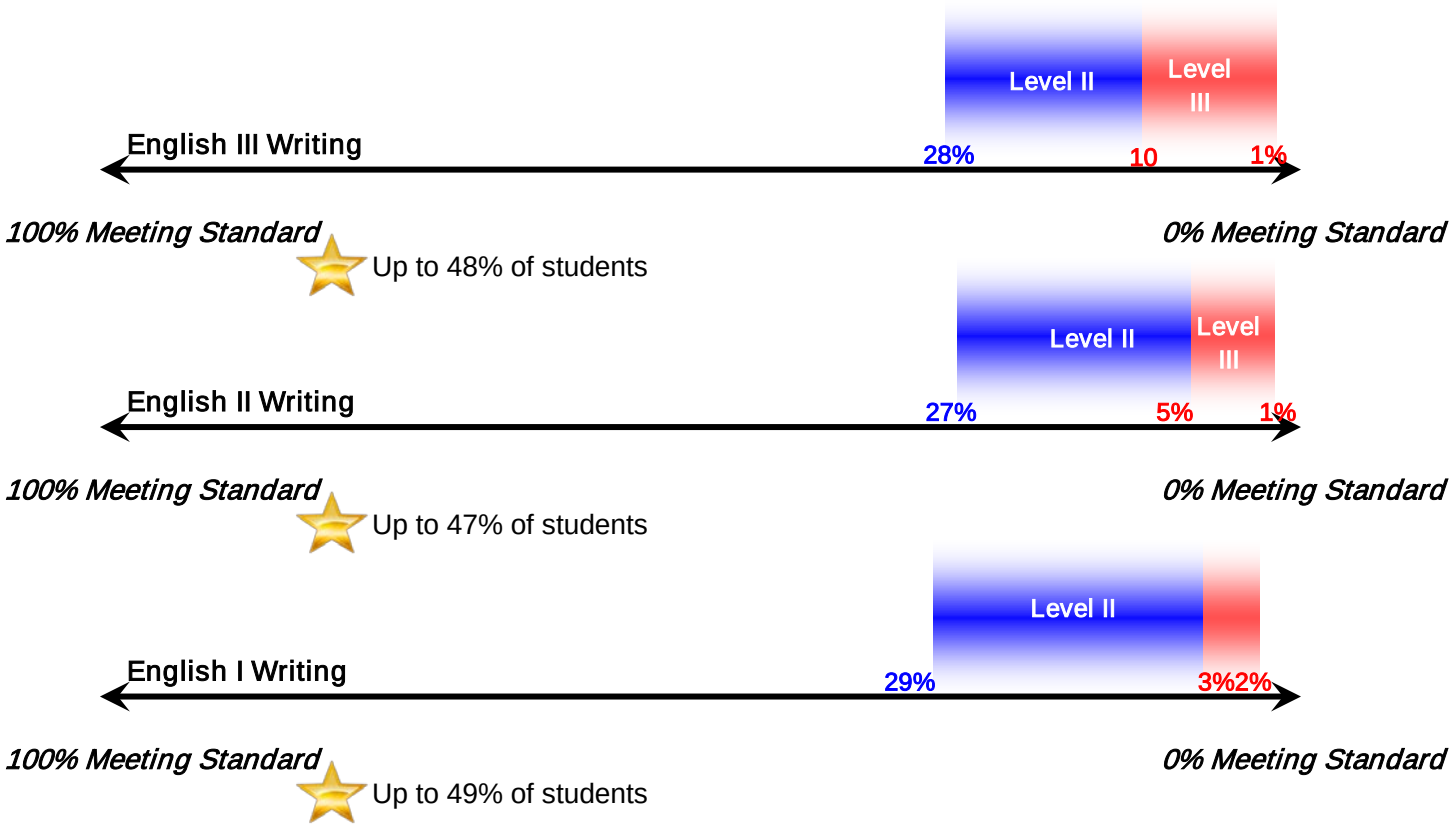
Option C — Additional Adjustment



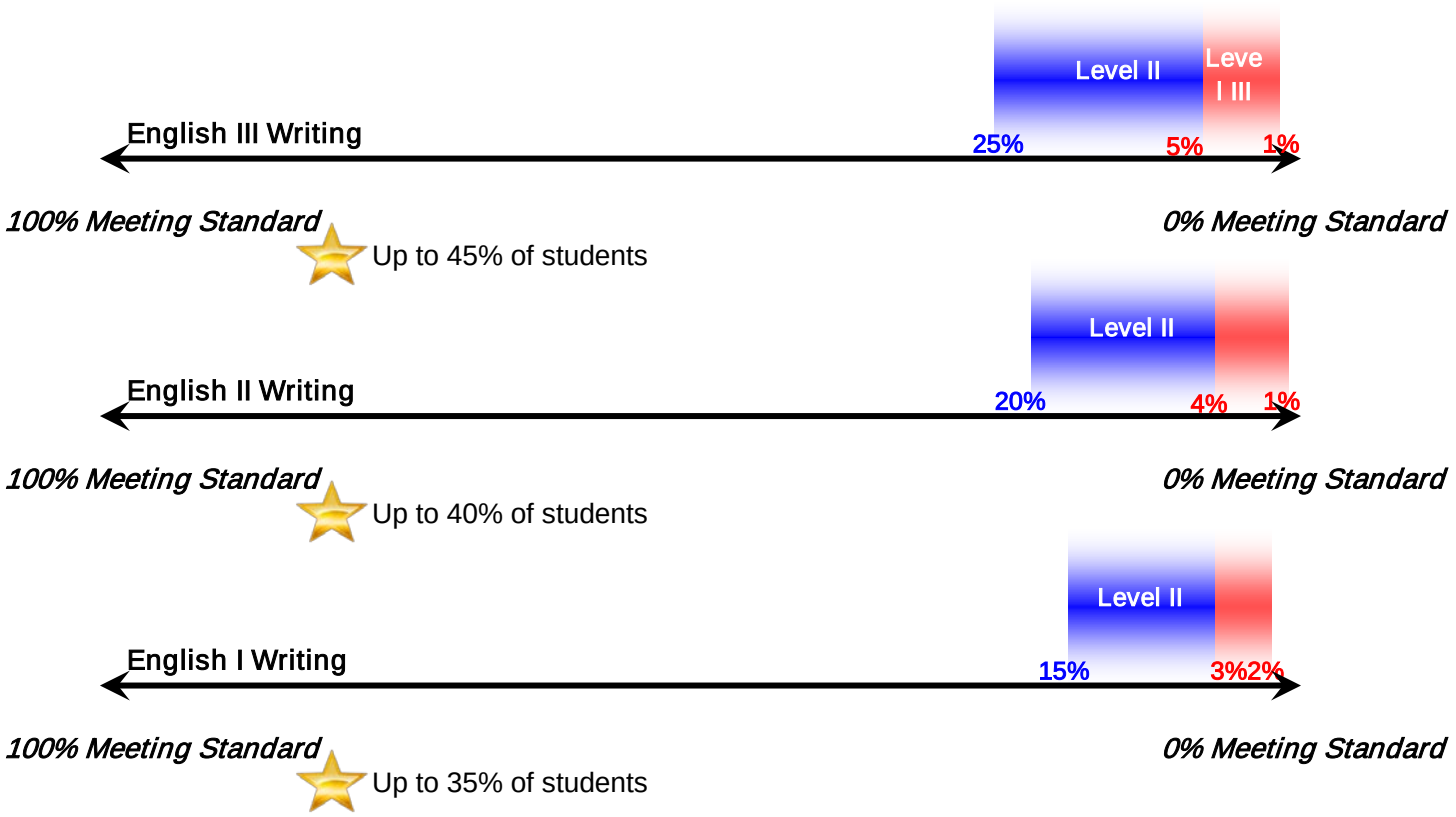
Option A — Based on Study Results



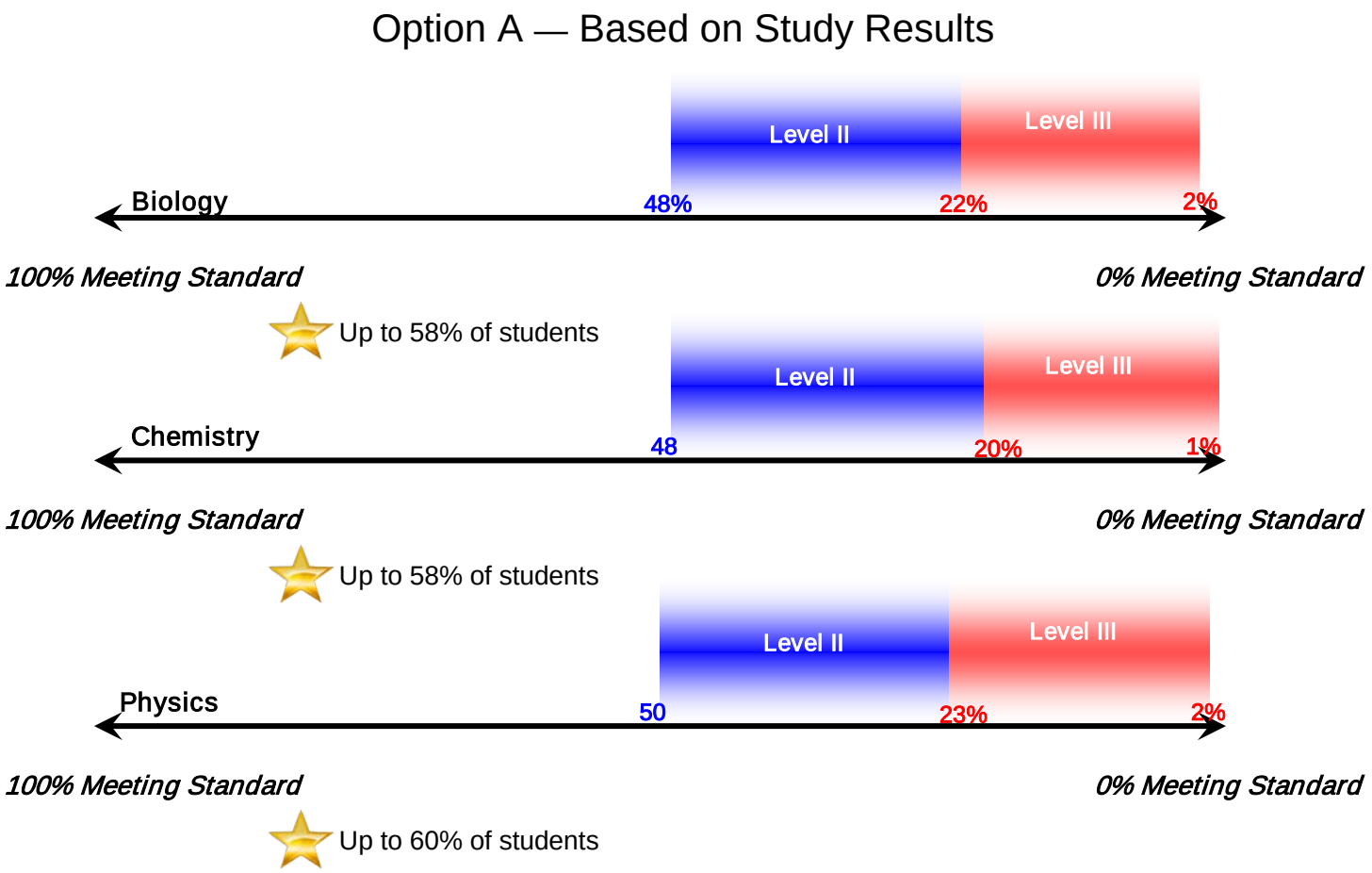
Option B — Adjusted for Motivation

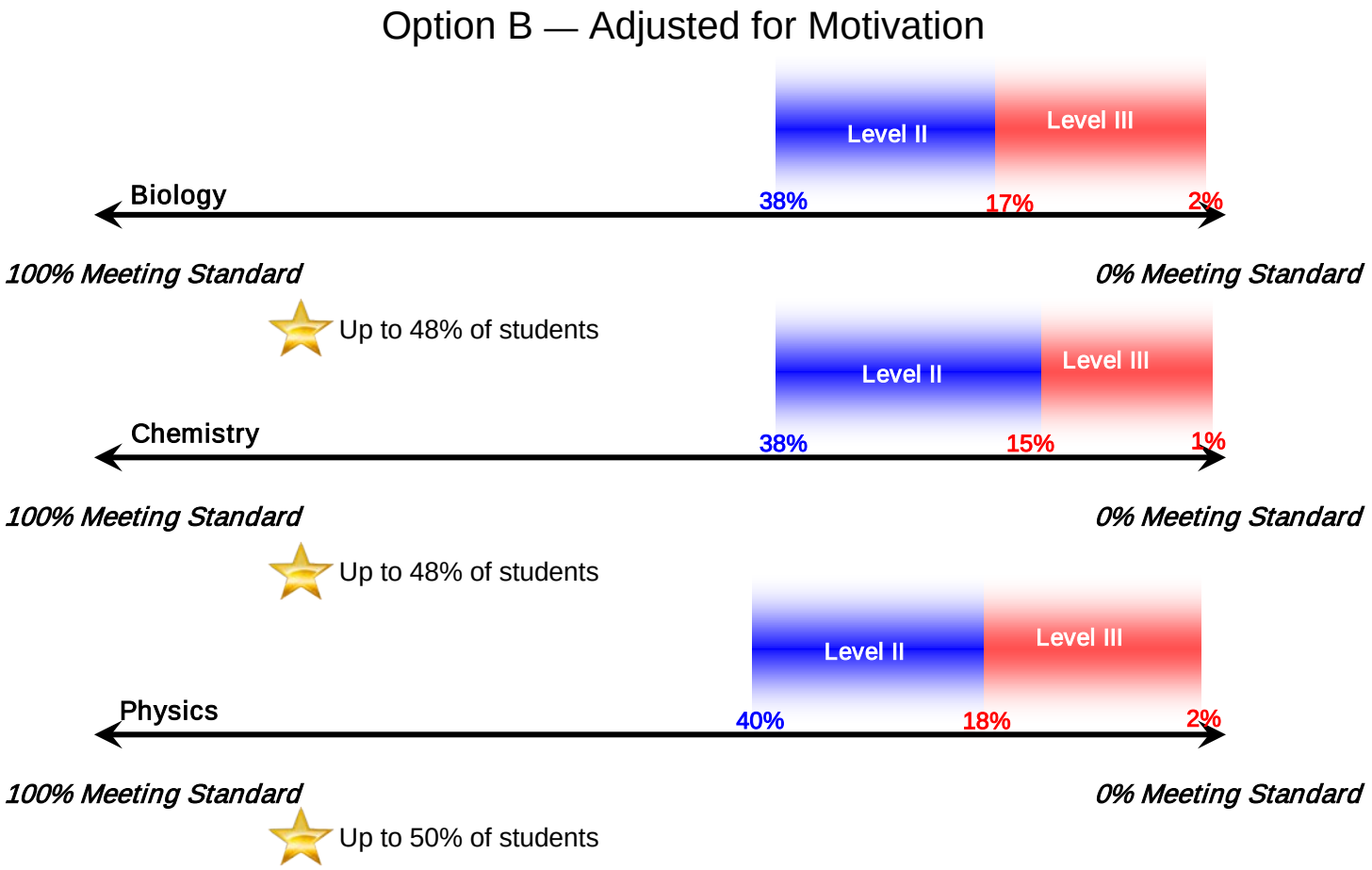


Option C — Additional Adjustment

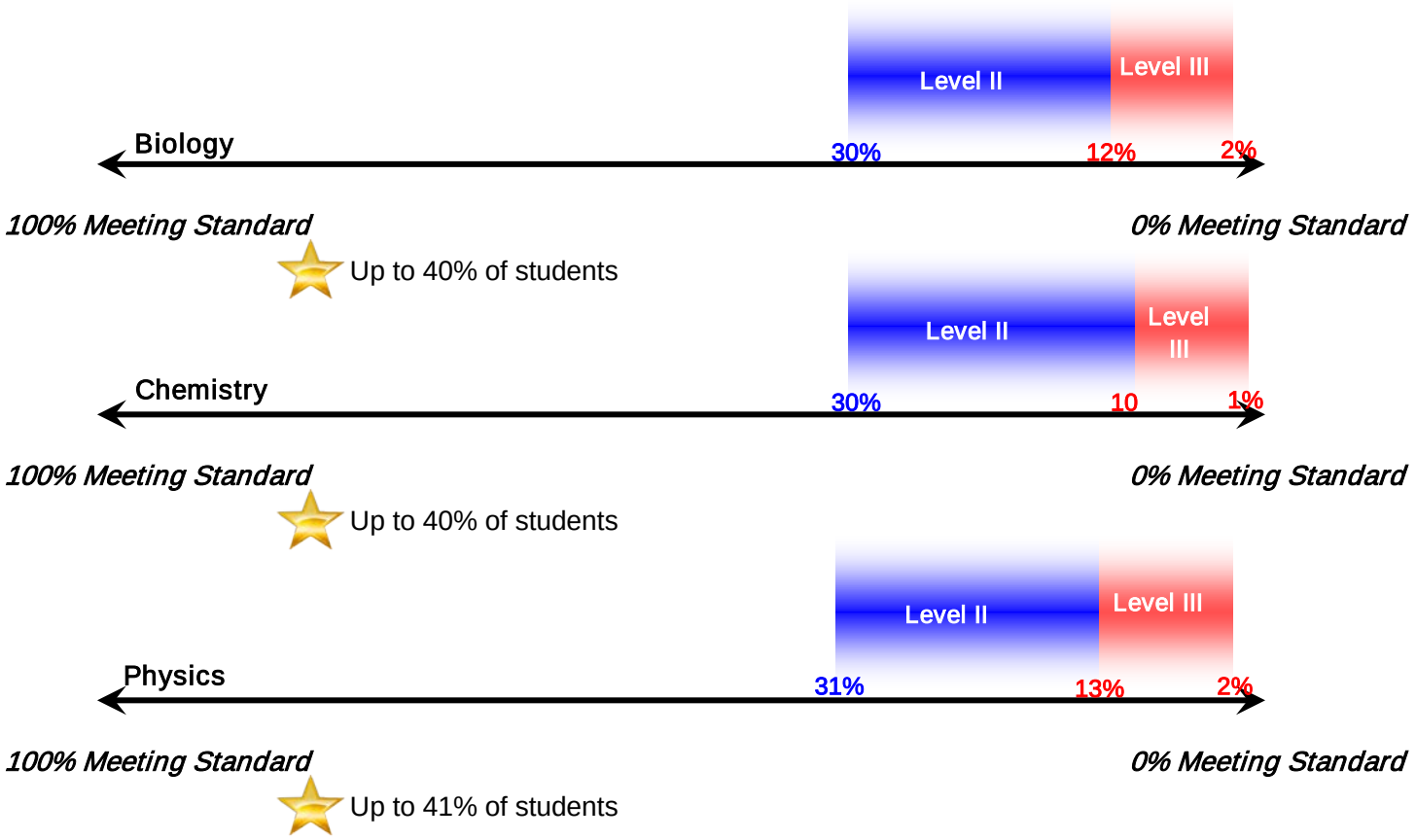


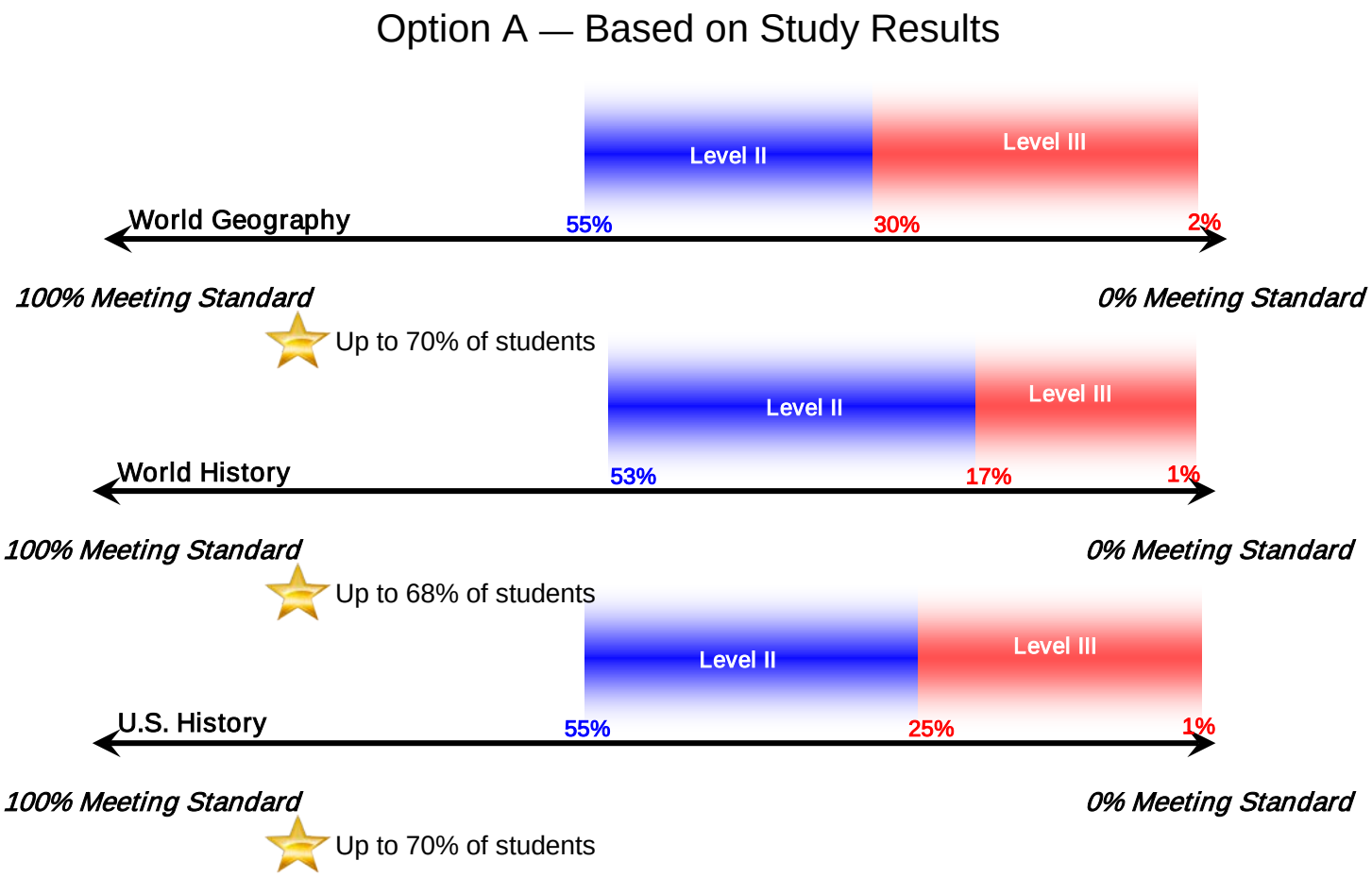
PART 3 – STAAR SCIENCE NEIGHBORHOOD OPTIONS



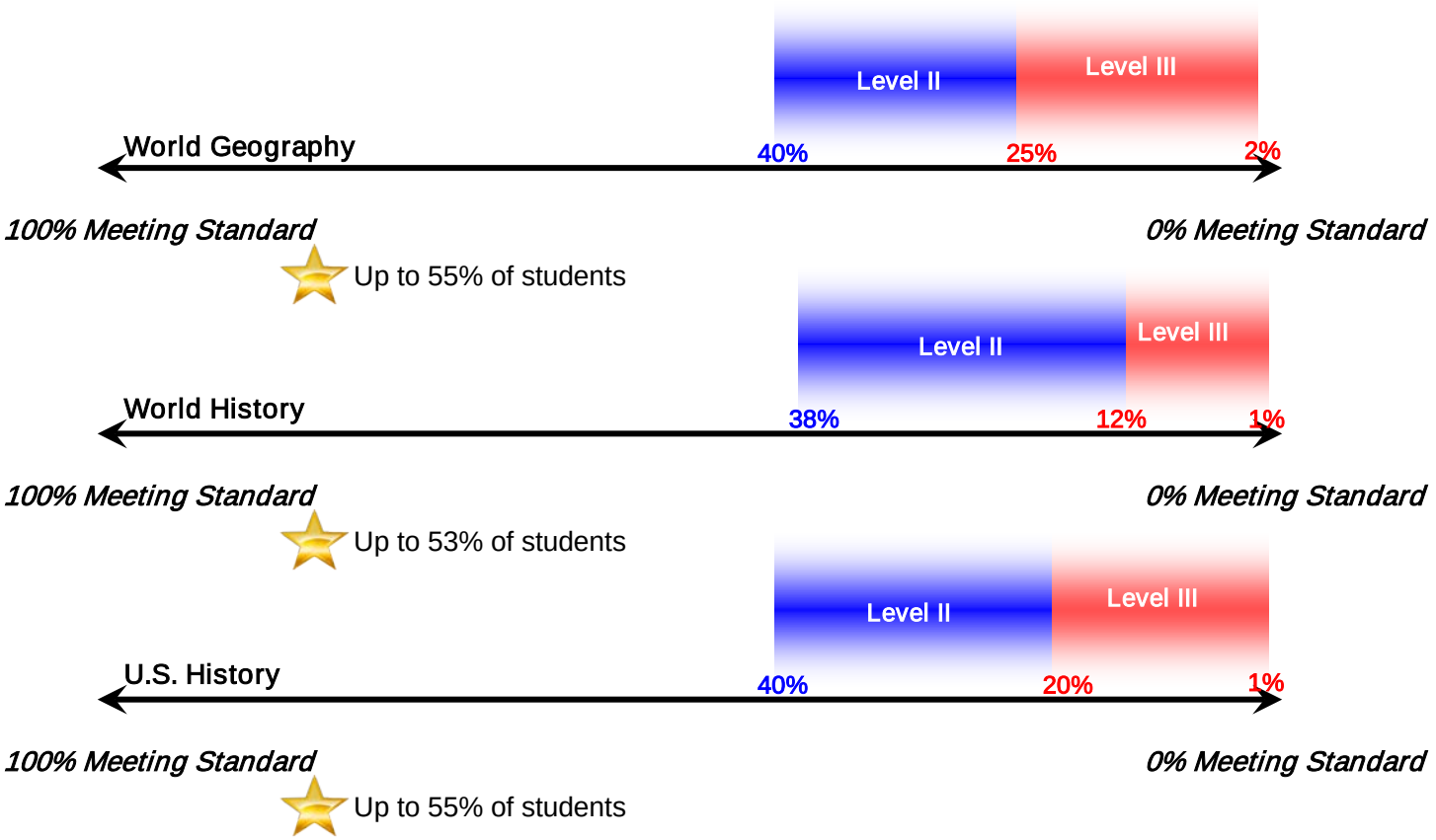

Option B — Adjusted for Motivation
Level II
Level III
Biology
38%
17%
2%
100% Meeting Standard
0% Meeting Standard
Up to 48% of students
Level II
Level III
Chemistry
38%
15%
1%
100% Meeting Standard
0% Meeting Standard
Up to 48% of students
Level II
Level III
Physics
40%
18%
2%
100% Meeting Standard
0% Meeting Standard
Up to 50% of students

Option C — Additional Adjustment

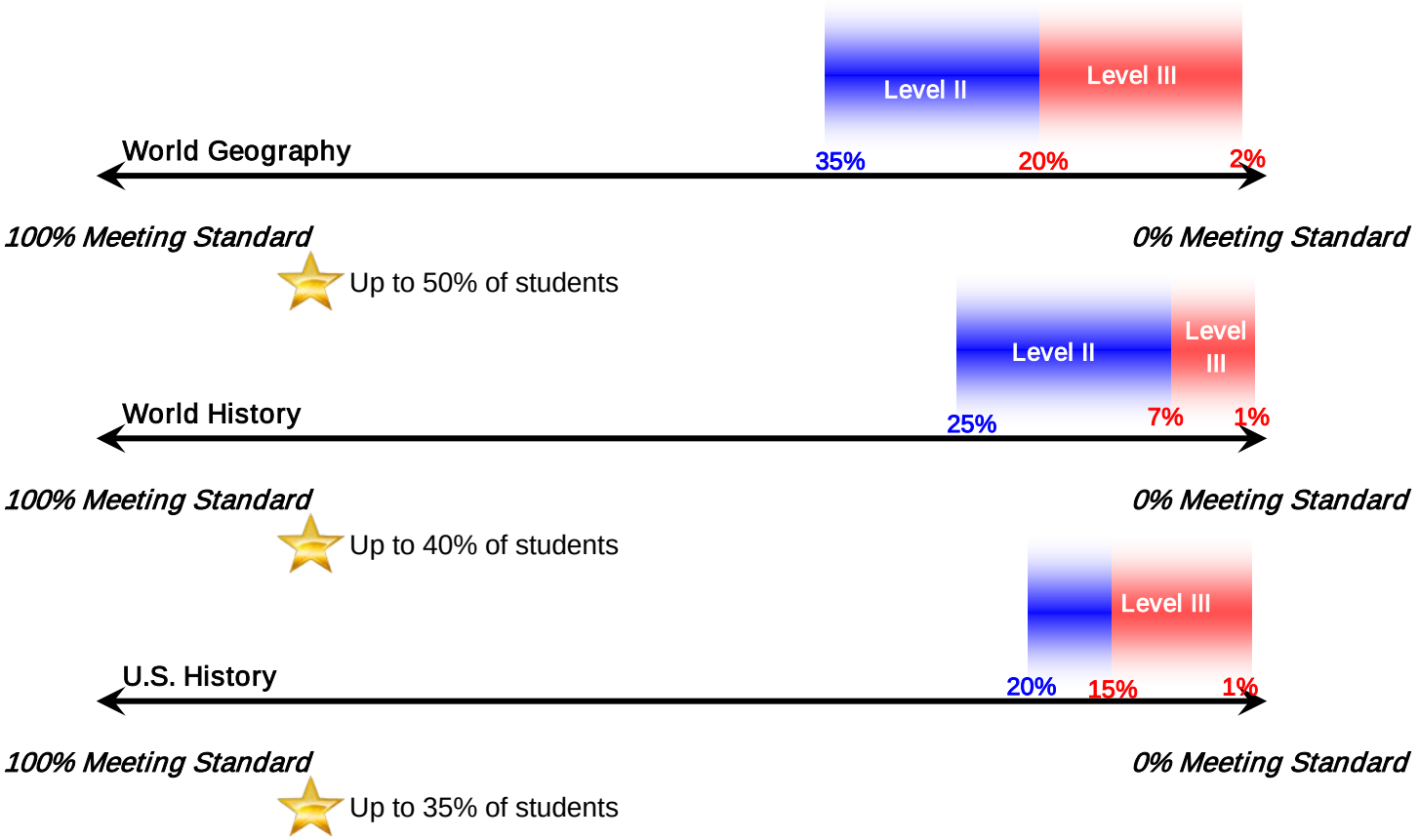



Option A — Based on Study Results
Level II
Level III
World Geography
55%
30%
2%
100% Meeting Standard
0% Meeting Standard
Up to 70% of students
Level II
Level III
World History
53%
17%
1%
100% Meeting Standard
0% Meeting Standard
Up to 68% of students
Level II
Level III
U.S. History
55%
25%
1%
100% Meeting Standard
0% Meeting Standard
Up to 70% of students

Option B — Adjusted for Motivation



Option C — Additional Adjustment



Appendix 9: Policy Committee Process Evaluation Summary

A total of 22 of the 28 policy committee members responded to the process evaluation survey. Their responses are summarized by section in the tables below.

SECTION 1: MEETING SUCCESS

Instructions: Check the column below that best reflects your opinion about the level of success of the various components of the meeting in which you have just participated. The activities were designed to help you both understand the process and be supportive of the recommendations made by the committee.

Summary of Reponses:

Meeting Component	Not Successful	Partially Successful	Successful	Very Successful	Omit
a. The purpose of the meeting	0%	23%	32%	41%	5%
b. Overview of the STAAR program and graduation plans	0%	14%	55%	27%	5%
c. Description of the performance labels and policy definitions	0%	9%	45%	45%	0%
d. Overview of the standard-setting process	0%	14%	45%	36%	5%
e. Overview of the validity studies	0%	27%	50%	23%	0%
f. Discussion of policy questions	5%	18%	50%	27%	0%
g. Presentation of neighborhood options with validity studies in Round 1	0%	23%	45%	32%	0%
h. Presentation of neighborhood options with validity studies in Round 2	0%	18%	55%	27%	0%
i. Presentation of neighborhood options with validity studies in Round 3	0%	23%	50%	23%	5%
j. Table discussions of neighborhood options during rounds	0%	5%	59%	36%	0%
k. Large group discussions of neighborhood options	0%	18%	41%	36%	5%
l. Presentation and discussions of feedback data	0%	14%	45%	41%	0%

SECTION 2: ADEQUACY OF TRAINING AND DISCUSSIONS

Instructions: How adequate were the following elements of the meeting?

Summary of Reponses:

Meeting Element	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
a. Amount of time spent training	0%	9%	41%	45%	5%
b. Feedback provided after table discussions	0%	9%	64%	23%	5%
c. Total amount of time for the tasks	0%	0%	41%	55%	5%

SECTION 3: OPPORTUNITY TO EXPRESS OPINIONS

Instructions: Did you have adequate opportunities during the session to express your professional opinions about the following elements?

Summary of Reponses:

Meeting Element	Not Adequate	Adequate	Omit
a. Answers to the policy questions	9%	86%	5%
b. Neighborhood options for STAAR EOC cut scores	5%	91%	5%

SECTION 4: SUPPORT AND INTERACTION DURING MEETING

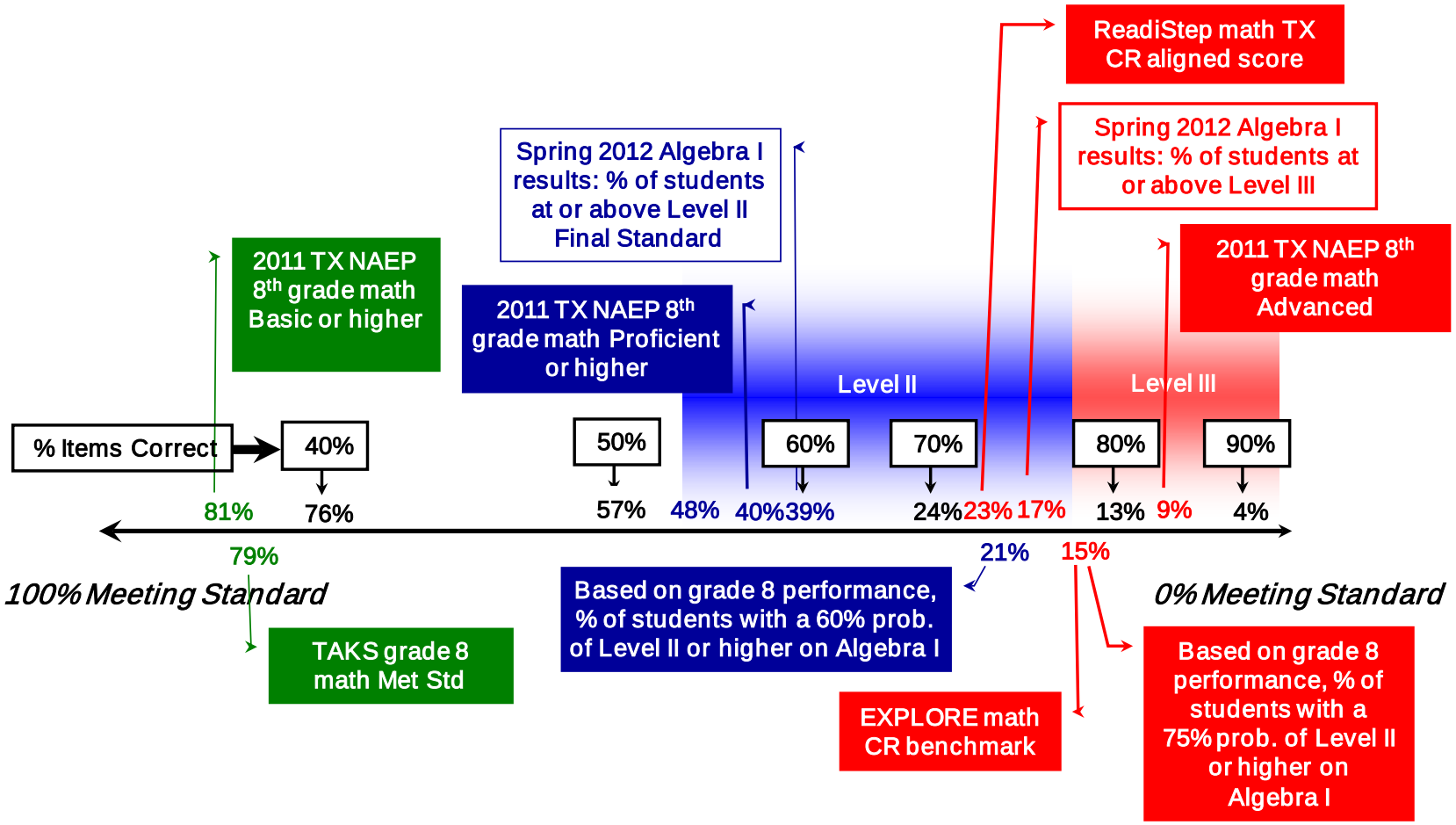
Instructions: Did you have adequate opportunities during the session regarding the following elements?

Summary of Reponses:

Meeting Element	Not Adequate	Adequate	Omit
a. Ask questions about the validity studies	5%	91%	5%
b. Ask questions about how the neighborhoods will be used	0%	95%	5%
c. Interact with your fellow committee members	0%	95%	5%

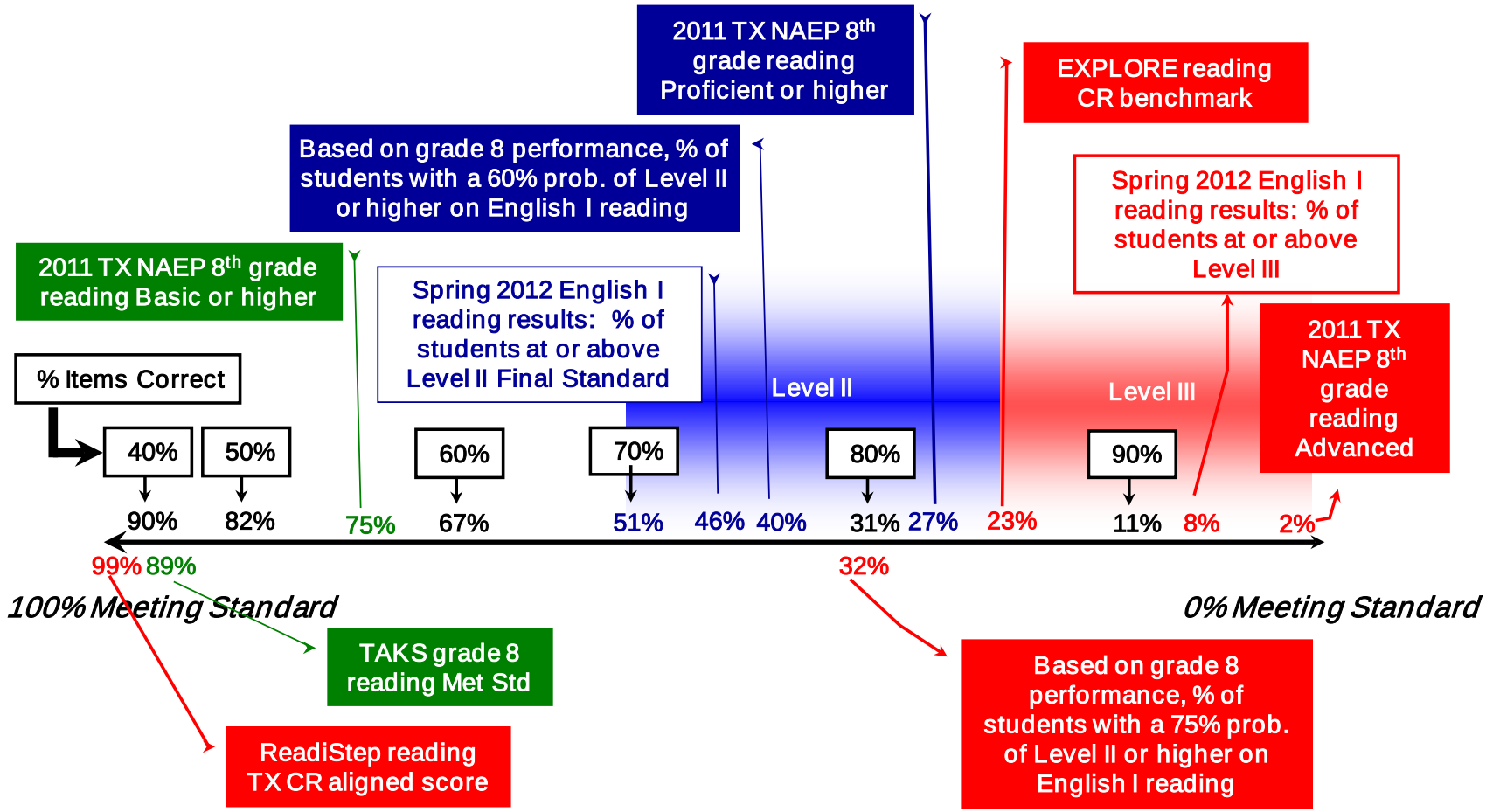
Appendix 10: STAAR Grade 8 Assessments and Grade 7 Writing Empirical Studies Number Lines

STAAR Grade 8 Mathematics



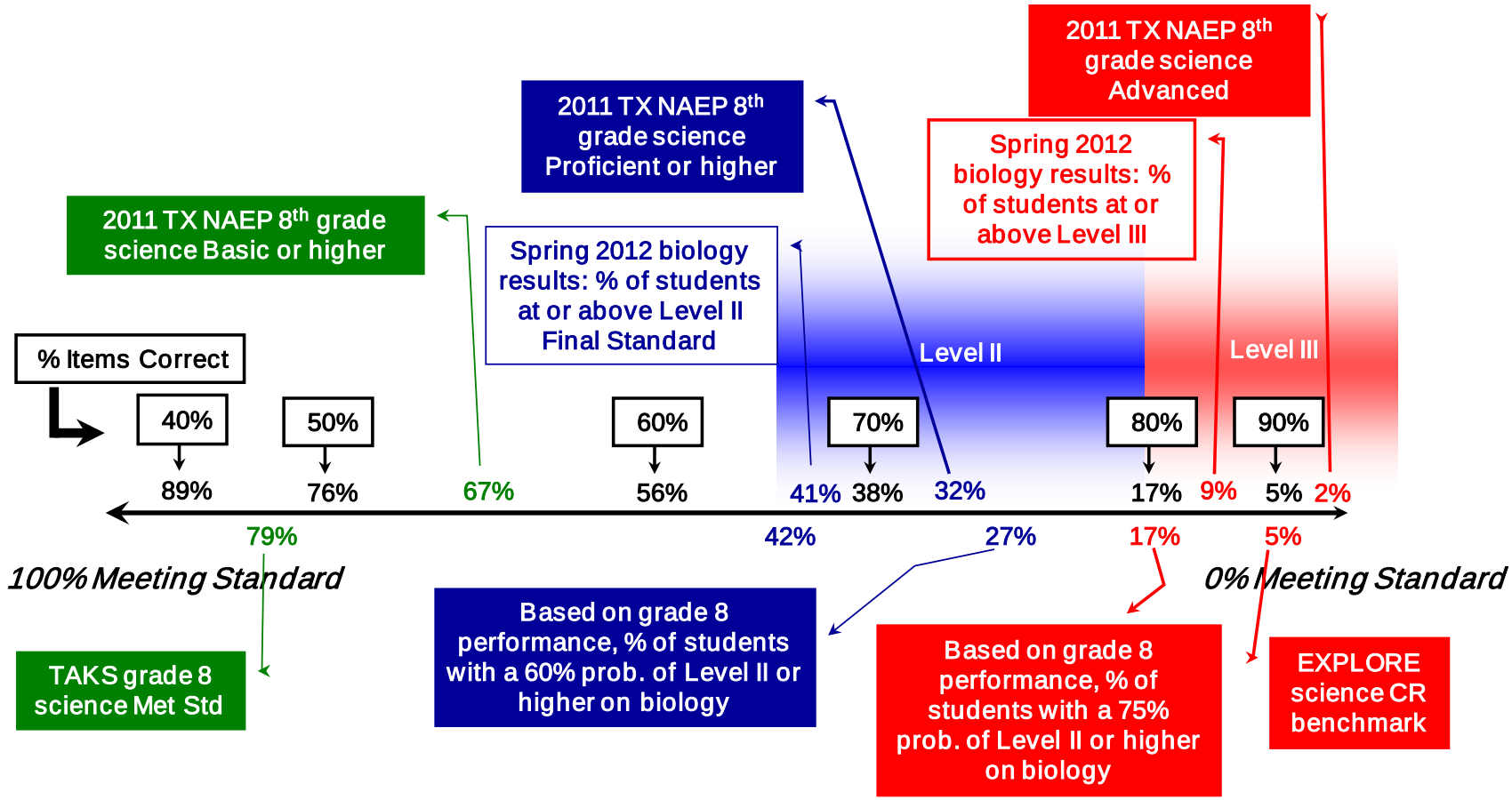
Abbreviation
CR = College Readiness

STAAR Grade 8 Reading



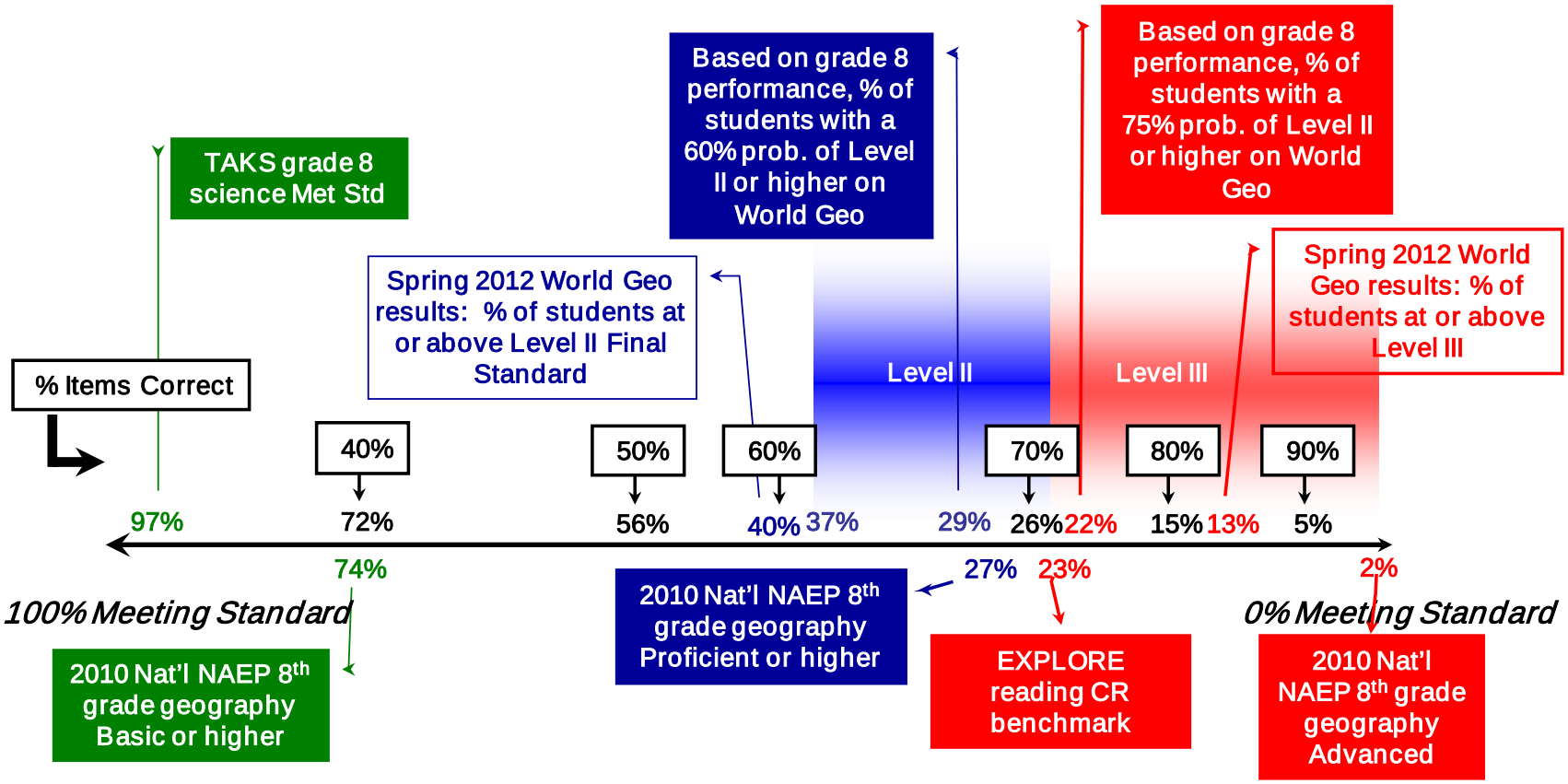
Abbreviation
CR = College Readiness

STAAR Grade 8 Science

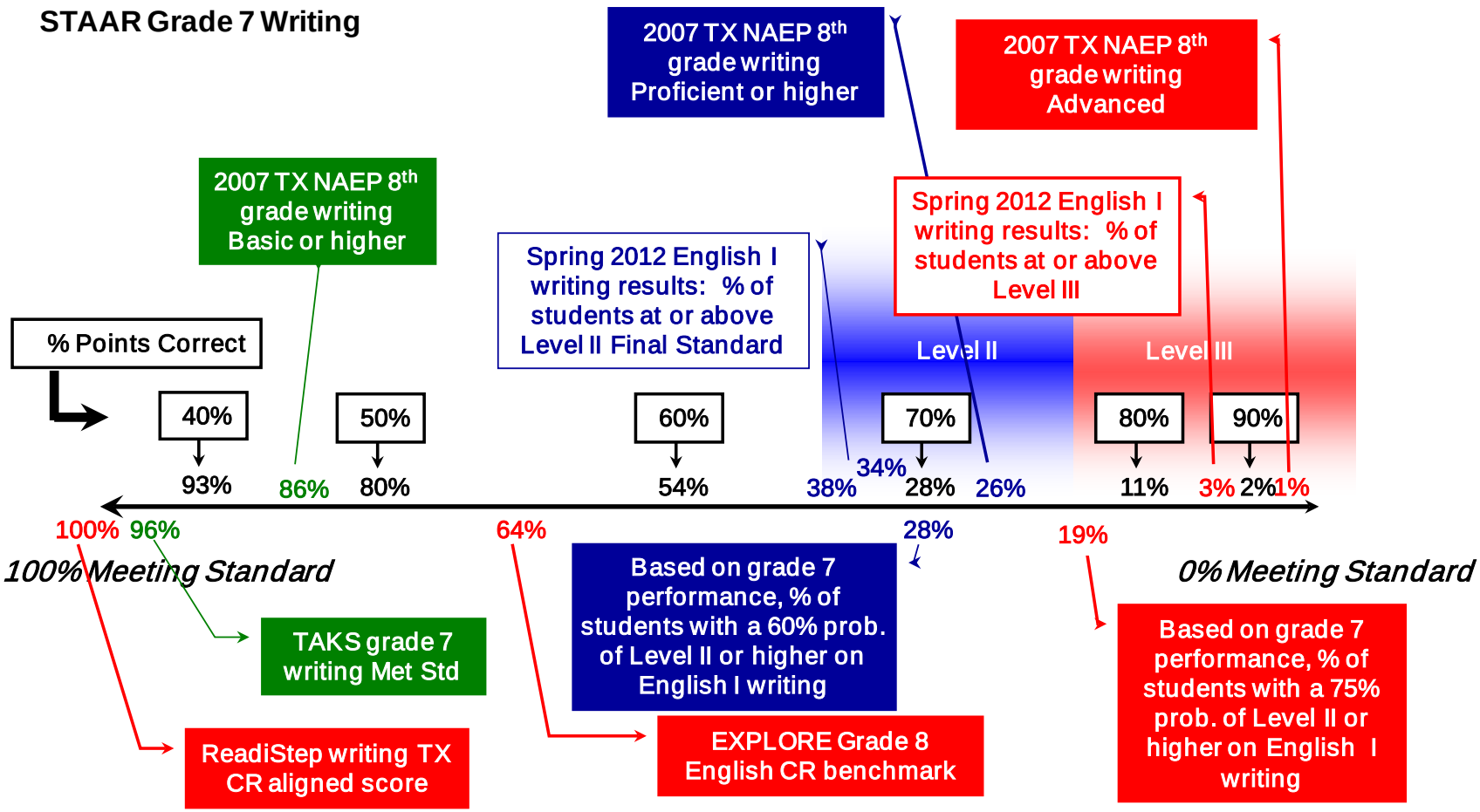


Abbreviation
CR = College Readiness

STAAR Grade 8 Social Studies



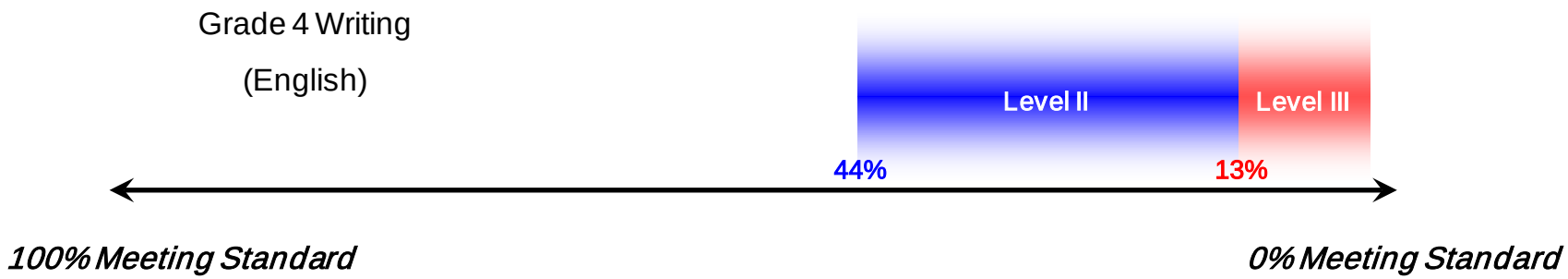
Abbreviation
CR = College Readiness

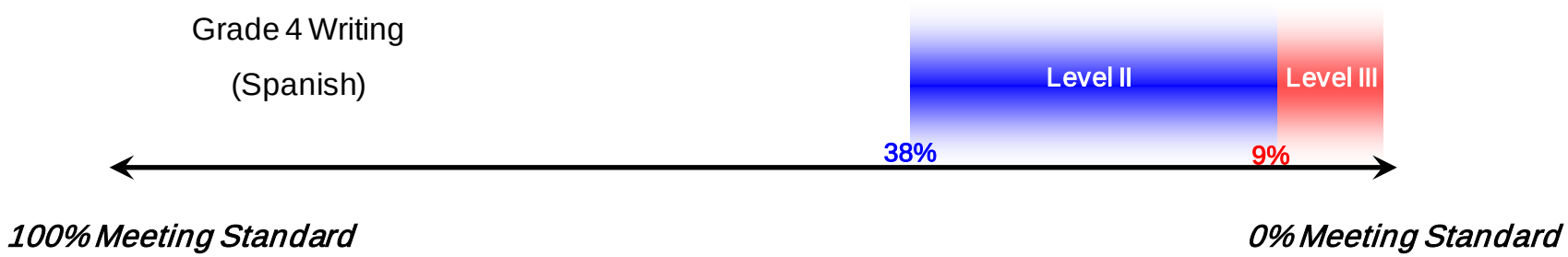


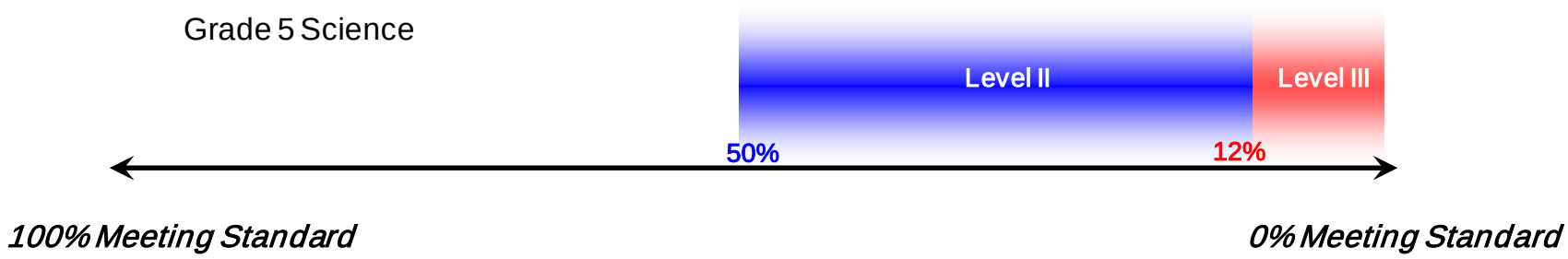
Abbreviation
CR = College Readiness

The grade 7 writing compositions are 44% of the total test score. The English I writing compositions are 52% of the total test score.

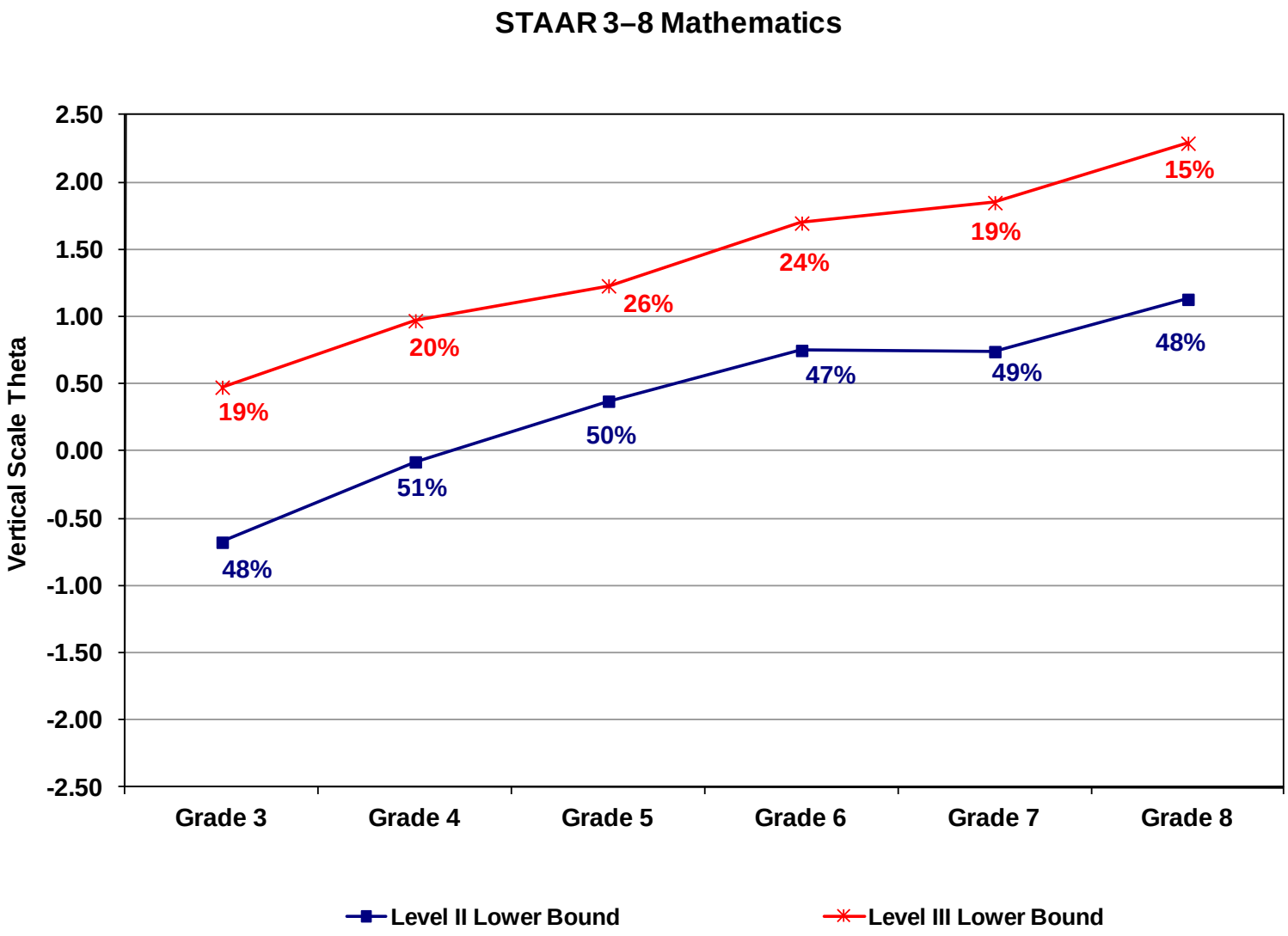
Appendix 11: STAAR Grade 4 English Writing, Grade 4 Spanish Writing, and Grade 5 Science Neighborhoods

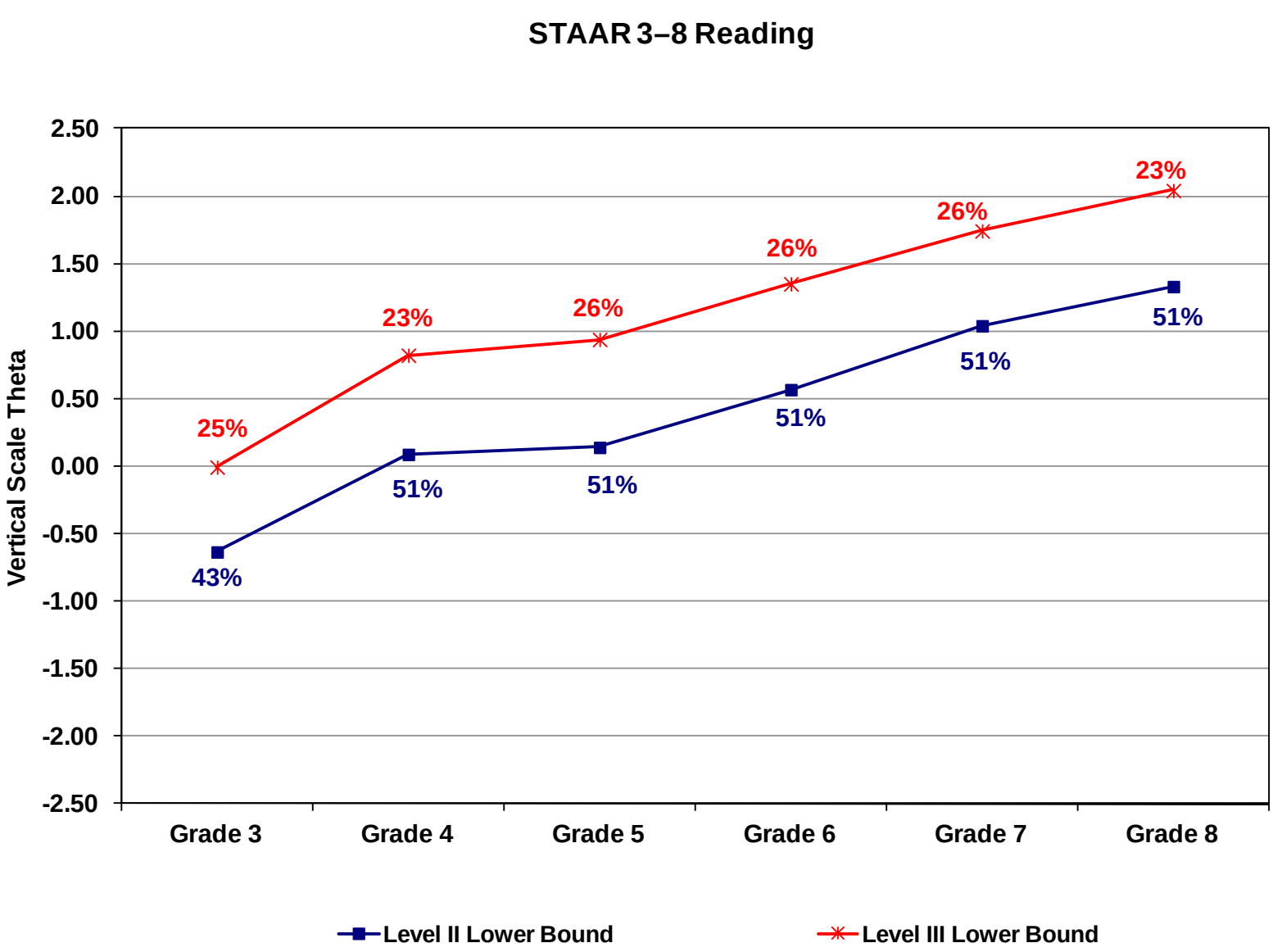


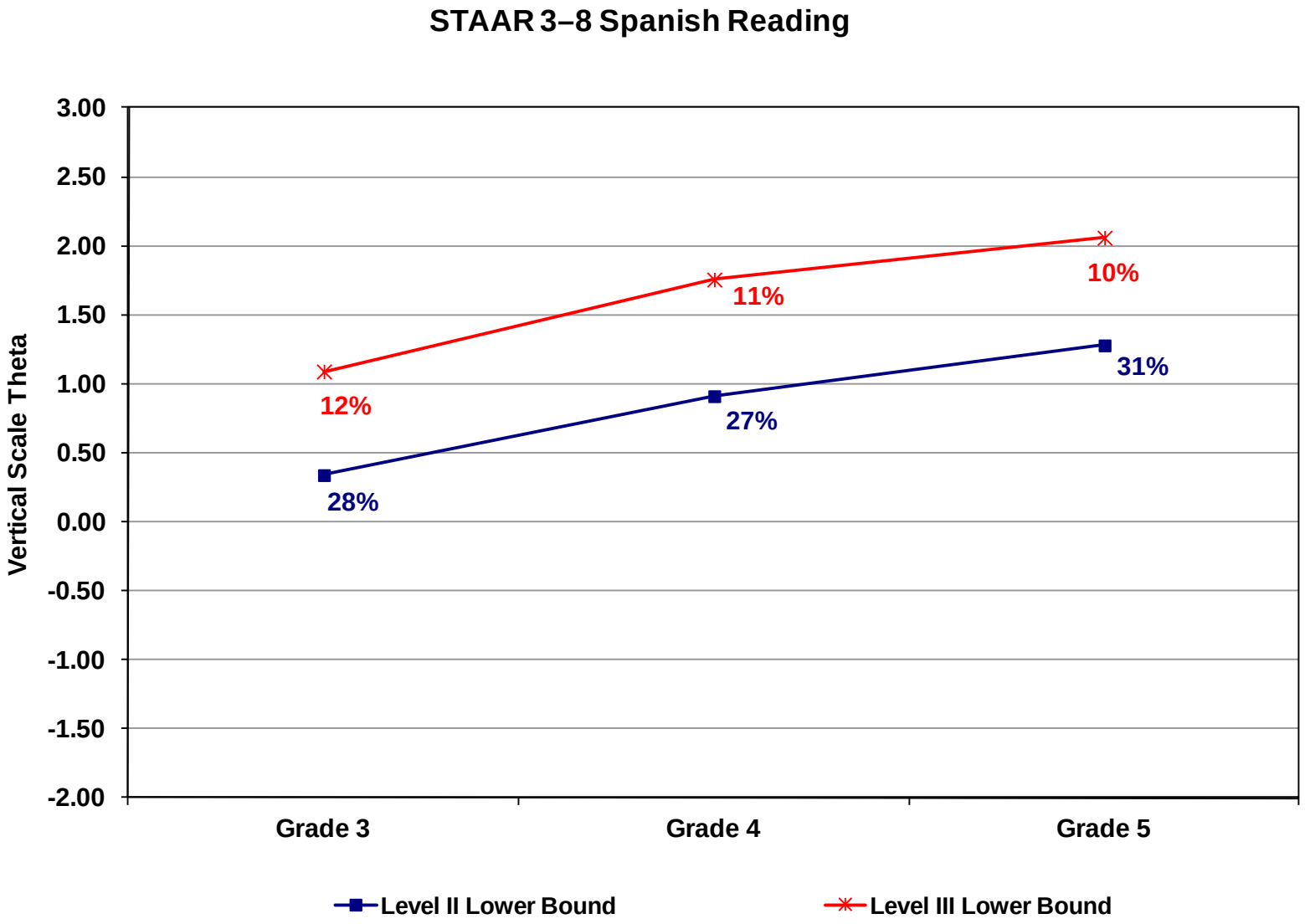




Appendix 12: STAAR 3–8 Mathematics and Reading Vertical Scale Neighborhoods




STAAR 3–8 Reading
Vertical Scale Theta
2.50
2.00
1.50
1.00
0.50
0.00
-0.50
-1.00
-1.50
-2.00
-2.50
Grade 3
Grade 4
Grade 5
Grade 6
Grade 7
Grade 8
25%
23%
26%
26%
26%
23%
43%
51%
51%
51%
51%
51%
Level II Lower Bound
Level III Lower Bound


STAAR 3–8 Spanish Reading
Vertical Scale Theta
3.00
2.50
2.00
1.50
1.00
0.50
0.00
-0.50
-1.00
-1.50
-2.00
Grade 3
Grade 4
Grade 5
12%
11%
10%
28%
27%
31%
Level II Lower Bound
Level III Lower Bound

Appendix 13: STAAR Standard-Setting Committee Composition

ENGLISH I, II, AND III READING COMMITTEE SUMMARY[1]

Current Position and Years of Experience in Education

		Years of Professional Experience in Education						
		Omit	1–5 years	6–10 years	11–15 years	16–20 years	More Than 20 years	Total
Current Position	Administrator	0	0	0	2	1	1	**5**
	Higher Education	0	0	0	0	0	1	**1**
	Teacher	0	0	3	0	2	2	**7**
	Other	0	0	0	0	0	0	**0**
	Total	**0**	**1**	**3**	**2**	**3**	**4**	**13**

Gender Distribution

Gender	N-Count
Female	9
Male	4

Ethnicity Distribution

Ethnicity	N-Count
African American	2
Asian or Pacific Islander	0
Hispanic	2
Multi-racial	1
Native American	1
White	7

Experience with Student Populations

Student Population	N-Count
General Education	12
Special Education	9
English Language Learners	10
Low Socioeconomic Status	11

District Type

Type	N-Count
Metro	2
Suburban	7
Rural	3
Did Not Respond	1

District Size

Type	N-Count
Large	9
Medium	2
Small	1
Did Not Respond	1

District Socioeconomic Status

Type	N-Count
High	1
Moderate	6
Low	5
Did Not Respond	1

[1] One panelist did not fill out the panelist information sheet and is not included in the numbers in this summary.

ENGLISH I, II, AND III WRITING COMMITTEE SUMMARY

Current Position and Years of Experience in Education

		Years of Professional Experience in Education						
		Omit	1–5 years	6–10 years	11–15 years	16–20 years	More Than 20 years	Total
Current Position	Administrator	0	0	0	1	1	2	**4**
	Higher Education	0	0	0	2	0	0	**2**
	Teacher	0	0	2	1	3	4	**10**
	Other	0	0	0	0	0	0	**0**
	Total	**0**	**0**	**2**	**4**	**4**	**6**	**16**

Gender Distribution

Gender	N-Count
Female	16
Male	0

Ethnicity Distribution

Ethnicity	N-Count
African American	1
Asian or Pacific Islander	0
Hispanic	4
Multi-racial	0
Native American	0
White	11

Experience with Student Populations

Student Population	N-Count
General Education	16
Special Education	11
English Language Learners	14
Low Socioeconomic Status	13

District Type

Type	N-Count
Metro	4
Suburban	5
Rural	5
Did Not Respond	2

District Size

Type	N-Count
Large	6
Medium	4
Small	4
Did Not Respond	2

District Socioeconomic Status

Type	N-Count
High	0
Moderate	6
Low	8
Did Not Respond	2

MATHEMATICS (ALGEBRA I, GEOMETRY, AND ALGEBRA II) COMMITTEE SUMMARY[2]

Current Position and Years of Experience in Education

		Years of Professional Experience in Education						
		Omit	1–5 years	6–10 years	11–15 years	16–20 years	More Than 20 years	Total
Current Position	Administrator	0	0	0	0	2	0	**2**
	Higher Education	0	1	0	0	1	1	**3**
	Teacher	0	1	3	2	5	8	**19**
	Other	0	0	0	0	0	1	**1**
	Total	**0**	**2**	**3**	**2**	**8**	**10**	**25**

Gender Distribution

Gender	N-Count
Female	11
Male	14

Ethnicity Distribution

Ethnicity	N-Count
African American	3
Asian or Pacific Islander	0
Hispanic	3
Multi-racial	0
Native American	0
White	19

Experience with Student Populations

Student Population	N-Count
General Education	22
Special Education	20
English Language Learners	20
Low Socioeconomic Status	22

District Type

Type	N-Count
Metro	6
Suburban	7
Rural	7
Did Not Respond	5

District Size

Type	N-Count
Large	5
Medium	11
Small	4
Did Not Respond	5

District Socioeconomic Status

Type	N-Count
High	1
Moderate	11
Low	8
Did Not Respond	5

[2] One panelist did not fill out the panelist information sheet and is not included in the numbers in this summary.

SCIENCE — BIOLOGY COMMITTEE SUMMARY

Current Position and Years of Experience in Education

		Years of Professional Experience in Education						
		Omit	1–5 years	6–10 years	11–15 years	16–20 years	More Than 20 years	Total
Current Position	Administrator	0	0	0	1	0	0	**1**
	Higher Education	0	0	0	0	0	1	**1**
	Teacher	0	1	4	2	0	5	**12**
	Other	0	0	0	1	0	1	**2**
	Total	**0**	**1**	**4**	**4**	**0**	**7**	**16**

Gender Distribution

Gender	N-Count
Female	12
Male	4

Ethnicity Distribution

Ethnicity	N-Count
African American	4
Asian or Pacific Islander	0
Hispanic	1
Multi-racial	0
Native American	0
White	11

Experience with Student Populations

Student Population	N-Count
General Education	15
Special Education	15
English Language Learners	13
Low Socioeconomic Status	15

District Type

Type	N-Count
Metro	3
Suburban	3
Rural	8
Did Not Respond	2

District Size

Type	N-Count
Large	3
Medium	7
Small	4
Did Not Respond	2

District Socioeconomic Status

Type	N-Count
High	1
Moderate	6
Low	7
Did Not Respond	2

SCIENCE — CHEMISTRY COMMITTEE SUMMARY

Current Position and Years of Experience in Education

		Years of Professional Experience in Education						
		Omit	1–5 years	6–10 years	11–15 years	16–20 years	More Than 20 years	Total
Current Position	Administrator	0	0	1	0	0	1	**2**
	Higher Education	0	0	0	0	0	0	**0**
	Teacher	1	1	2	1	2	3	**10**
	Other	1	1	0	0	1	0	**2**
	Total	**2**	**1**	**3**	**1**	**3**	**4**	**14**

Gender Distribution

Gender	N-Count
Female	8
Male	6

Ethnicity Distribution

Ethnicity	N-Count
African American	2
Asian or Pacific Islander	0
Hispanic	5
Multi-racial	0
Native American	0
White	7

Experience with Student Populations

Student Population	N-Count
General Education	13
Special Education	11
English Language Learners	12
Low Socioeconomic Status	12

District Type

Type	N-Count
Metro	4
Suburban	5
Rural	2
Did Not Respond	3

District Size

Type	N-Count
Large	7
Medium	2
Small	2
Did Not Respond	3

District Socioeconomic Status

Type	N-Count
High	1
Moderate	4
Low	6
Did Not Respond	3

SCIENCE — PHYSICS COMMITTEE SUMMARY

Current Position and Years of Experience in Education

		Years of Professional Experience in Education						
		Omit	1–5 years	6–10 years	11–15 years	16–20 years	More Than 20 years	Total
Current Position	Administrator	0	0	0	0	0	0	**0**
	Higher Education	0	0	0	0	0	1	**1**
	Teacher	0	1	2	2	1	7	**13**
	Other	0	0	0	0	1	0	**1**
	Total	**0**	**1**	**2**	**2**	**2**	**8**	**15**

Gender Distribution

Gender	N-Count
Female	11
Male	4

Ethnicity Distribution

Ethnicity	N-Count
African American	1
Asian or Pacific Islander	2
Hispanic	3
Multi-racial	0
Native American	0
White	9

Experience with Student Populations

Student Population	N-Count
General Education	13
Special Education	10
English Language Learners	9
Low Socioeconomic Status	13

District Type

Type	N-Count
Metro	5
Suburban	2
Rural	8
Did Not Respond	0

District Size

Type	N-Count
Large	4
Medium	6
Small	5
Did Not Respond	0

District Socioeconomic Status

Type	N-Count
High	0
Moderate	5
Low	9
Did Not Respond	1

SOCIAL STUDIES — WORLD GEOGRAPHY COMMITTEE SUMMARY

Current Position and Years of Experience in Education

		Years of Professional Experience in Education						
		Omit	1–5 years	6–10 years	11–15 years	16–20 years	More Than 20 years	Total
Current Position	Administrator	0	0	0	0	0	1	**1**
	Higher Education	0	0	0	0	0	2	**2**
	Teacher	0	1	2	3	1	2	**9**
	Other	0	0	0	1	0	1	**2**
	Total	**0**	**1**	**2**	**4**	**1**	**6**	**14**

Gender Distribution

Gender	N-Count
Female	6
Male	8

Ethnicity Distribution

Ethnicity	N-Count
African American	3
Asian or Pacific Islander	0
Hispanic	2
Multi-racial	0
Native American	0
White	9

Experience with Student Populations

Student Population	N-Count
General Education	12
Special Education	11
English Language Learners	12
Low Socioeconomic Status	11

District Type

Type	N-Count
Metro	1
Suburban	5
Rural	5
Did Not Respond	3

District Size

Type	N-Count
Large	4
Medium	4
Small	3
Did Not Respond	3

District Socioeconomic Status

Type	N-Count
High	0
Moderate	6
Low	5
Did Not Respond	3

SOCIAL STUDIES — WORLD HISTORY COMMITTEE SUMMARY

Current Position and Years of Experience in Education

		Years of Professional Experience in Education						
		Omit	1–5 years	6–10 years	11–15 years	16–20 years	More Than 20 years	Total
Current Position	Administrator	0	0	0	0	0	0	**0**
	Higher Education	0	0	0	1	0	0	**1**
	Teacher	0	3	2	5	2	0	**12**
	Other	0	0	0	0	0	0	**0**
	Total	**0**	**3**	**2**	**6**	**2**	**0**	**13**

Gender Distribution

Gender	N-Count
Female	6
Male	7

Ethnicity Distribution

Ethnicity	N-Count
African American	0
Asian or Pacific Islander	1
Hispanic	3
Multi-racial	2
Native American	0
White	7

Experience with Student Populations

Student Population	N-Count
General Education	12
Special Education	9
English Language Learners	9
Low Socioeconomic Status	9

District Type

Type	N-Count
Metro	3
Suburban	5
Rural	3
Did Not Respond	2

District Size

Type	N-Count
Large	4
Medium	4
Small	3
Did Not Respond	2

District Socioeconomic Status

Type	N-Count
High	0
Moderate	7
Low	4
Did Not Respond	2

SOCIAL STUDIES — U.S. HISTORY COMMITTEE SUMMARY

Current Position and Years of Experience in Education

		Years of Professional Experience in Education						
		Omit	1–5 years	6–10 years	11–15 years	16–20 years	More Than 20 years	Total
Current Position	Administrator	0	0	0	0	1	2	**3**
	Higher Education	0	0	0	0	0	0	**0**
	Teacher	0	1	1	1	2	3	**8**
	Other	0	0	0	0	0	1	**1**
	Total	**0**	**1**	**1**	**1**	**3**	**6**	**12**

Gender Distribution

Gender	N-Count
Female	8
Male	4

Ethnicity Distribution

Ethnicity	N-Count
African American	0
Asian or Pacific Islander	0
Hispanic	2
Multi-racial	1
Native American	0
White	9

Experience with Student Populations

Student Population	N-Count
General Education	10
Special Education	10
English Language Learners	8
Low Socioeconomic Status	11

District Type

Type	N-Count
Metro	5
Suburban	4
Rural	3
Did Not Respond	0

District Size

Type	N-Count
Large	5
Medium	6
Small	1
Did Not Respond	0

District Socioeconomic Status

Type	N-Count
High	0
Moderate	4
Low	8
Did Not Respond	0

GRADE 8 MATHEMATICS

Current Position and Years of Experience in Education

		Years of Professional Experience in Education						
		Omit	1–5 years	6–10 years	11–15 years	16–20 years	More Than 20 years	Total
Current Position	Administrator	0	0	0	0	0	2	**2**
	Teacher	0	0	4	2	1	2	**9**
	Other	0	0	0	0	1	1	**2**
	Total	**0**	**0**	**4**	**2**	**2**	**5**	**13**

Gender Distribution

Gender	N-Count
Female	7
Male	6

Ethnicity Distribution

Ethnicity	N-Count
African American	0
Asian or Pacific Islander	1
Hispanic	0
Multi-racial	1
Native American	0
White	11

Experience with Student Populations

Student Population	N-Count
General Education	13
Special Education	12
English Language Learners	10
Low Socioeconomic Status	12

District Type

Type	N-Count
Metro	3
Suburban	7
Rural	2
Other	0
Did Not Respond	1

District Size

Type	N-Count
Large	6
Medium	4
Small	2
Other	0
Did Not Respond	1

District Socioeconomic Status

Type	N-Count
High	0
Moderate	7
Low	5
Other	0
Did Not Respond	1

GRADE 8 READING

Current Position and Years of Experience in Education

		Years of Professional Experience in Education						
		Omit	1–5 years	6–10 years	11–15 years	16–20 years	More Than 20 years	Total
Current Position	Administrator	0	0	0	0	1	0	**1**
	Teacher	1	0	2	3	3	3	**12**
	Other	0	0	0	0	1	0	**1**
	Total	**1**	**0**	**2**	**3**	**5**	**3**	**14**

Gender Distribution

Gender	N-Count
Female	12
Male	2

Ethnicity Distribution

Ethnicity	N-Count
African American	0
Asian or Pacific Islander	0
Hispanic	3
Multi-racial	1
Native American	1
White	9

Experience with Student Populations

Student Population	N-Count
General Education	14
Special Education	12
English Language Learners	13
Low Socioeconomic Status	13

District Type

Type	N-Count
Metro	3
Suburban	3
Rural	8
Other	0
Did Not Respond	0

District Size

Type	N-Count
Large	4
Medium	4
Small	5
Other	1
Did Not Respond	0

District Socioeconomic Status

Type	N-Count
High	0
Moderate	6
Low	7
Other	1
Did Not Respond	0

GRADE 8 SCIENCE

Current Position and Years of Experience in Education

		Years of Professional Experience in Education						
		Omit	1–5 years	6–10 years	11–15 years	16–20 years	More Than 20 years	Total
Current Position	Administrator	0	0	0	0	1	1	**2**
	Teacher	0	0	2	5	0	3	**10**
	Other	0	0	0	0	0	0	**0**
	Total	**0**	**0**	**2**	**5**	**1**	**4**	**12**

Gender Distribution

Gender	N-Count
Female	10
Male	2

Ethnicity Distribution

Ethnicity	N-Count
African American	2
Asian or Pacific Islander	0
Hispanic	0
Multi-racial	3
Native American	1
White	6

Experience with Student Populations

Student Population	N-Count
General Education	12
Special Education	12
English Language Learners	10
Low Socioeconomic Status	12

District Type

Type	N-Count
Metro	1
Suburban	3
Rural	7
Other	0
Did Not Respond	1

District Size

Type	N-Count
Large	2
Medium	6
Small	3
Other	0
Did Not Respond	1

District Socioeconomic Status

Type	N-Count
High	0
Moderate	4
Low	7
Other	0
Did Not Respond	1

GRADE 8 SOCIAL STUDIES

Current Position and Years of Experience in Education

		Years of Professional Experience in Education						
		Omit	1–5 years	6–10 years	11–15 years	16–20 years	More Than 20 years	Total
Current Position	Administrator	0	0	0	0	0	1	**1**
	Teacher	0	0	3	5	2	1	**11**
	Other	0	0	0	0	0	0	**0**
	Total	**0**	**0**	**3**	**5**	**2**	**2**	**12**

Gender Distribution

Gender	N-Count
Female	8
Male	4

Ethnicity Distribution

Ethnicity	N-Count
African American	1
Asian or Pacific Islander	0
Hispanic	1
Multi-racial	2
Native American	0
White	8

Experience with Student Populations

Student Population	N-Count
General Education	11
Special Education	12
English Language Learners	10
Low Socioeconomic Status	11

District Type

Type	N-Count
Metro	2
Suburban	5
Rural	5
Other	0
Did Not Respond	0

District Size

Type	N-Count
Large	4
Medium	5
Small	3
Other	0
Did Not Respond	0

District Socioeconomic Status

Type	N-Count
High	2
Moderate	1
Low	8
Other	1
Did Not Respond	0

GRADE 7 WRITING (OCTOBER 2012)

Current Position and Years of Experience in Education

		Years of Professional Experience in Education						
		Omit	1–5 years	6–10 years	11–15 years	16–20 years	More Than 20 years	Total
Current Position	Administrator	0	0	0	2	1	1	**4**
	Teacher	0	0	1	0	4	3	**8**
	Other	0	0	1	0	0	0	**1**
	Total	**0**	**0**	**2**	**2**	**5**	**4**	**13**

Gender Distribution

Gender	N-Count
Female	11
Male	2

Ethnicity Distribution

Ethnicity	N-Count
African American	2
Asian or Pacific Islander	0
Hispanic	1
Multi-racial	3
Native American	0
White	7

Experience with Student Populations

Student Population	N-Count
General Education	13
Special Education	13
English Language Learners	12
Low Socioeconomic Status	13

District Type

Type	N-Count
Metro	4
Suburban	5
Rural	3
Other	1
Did Not Respond	0

District Size

Type	N-Count
Large	6
Medium	5
Small	2
Other	0
Did Not Respond	0

District Socioeconomic Status

Type	N-Count
High	1
Moderate	6
Low	5
Other	1
Did Not Respond	0

GRADE 7 WRITING (NOVEMBER 2012)

Current Position and Years of Experience in Education

		Years of Professional Experience in Education						
		Omit	1–5 years	6–10 years	11–15 years	16–20 years	More Than 20 years	Total
Current Position	Administrator	0	0	0	2	0	1	**3**
	Teacher	0	0	1	0	3	2	**6**
	Other	0	0	1	0	0	0	**1**
	Total	**0**	**0**	**2**	**2**	**3**	**3**	**10**

Gender Distribution

Gender	N-Count
Female	8
Male	2

Ethnicity Distribution

Ethnicity	N-Count
African American	2
Asian or Pacific Islander	0
Hispanic	1
Multi-racial	3
Native American	0
White	4

Experience with Student Populations

Student Population	N-Count
General Education	10
Special Education	10
English Language Learners	10
Low Socioeconomic Status	10

District Type

Type	N-Count
Metro	3
Suburban	4
Rural	2
Other	1
Did Not Respond	0

District Size

Type	N-Count
Large	6
Medium	2
Small	2
Other	0
Did Not Respond	0

District Socioeconomic Status

Type	N-Count
High	1
Moderate	6
Low	3
Other	0
Did Not Respond	0

GRADES 6 AND 7 MATHEMATICS

Current Position and Years of Experience in Education

		Years of Professional Experience in Education						
		Omit	1–5 years	6–10 years	11–15 years	16–20 years	More Than 20 years	Total
Current Position	Administrator	0	0	1	0	0	0	**1**
	Teacher	0	3	5	1	0	3	**12**
	Other	0	0	1	0	0	1	**2**
	Total	**0**	**3**	**7**	**1**	**0**	**4**	**15**

Gender Distribution

Gender	N-Count
Female	12
Male	3

Ethnicity Distribution

Ethnicity	N-Count
African American	2
Asian or Pacific Islander	0
Hispanic	0
Multi-racial	2
Native American	0
White	10
Did Not Respond	1

Experience with Student Populations

Student Population	N-Count
General Education	14
Special Education	14
English Language Learners	15
Low Socioeconomic Status	15

District Type

Type	N-Count
Metro	4
Suburban	4
Rural	6
Other	0
Did Not Respond	1

District Size

Type	N-Count
Large	4
Medium	6
Small	4
Other	0
Did Not Respond	1

District Socioeconomic Status

Type	N-Count
High	0
Moderate	5
Low	8
Other	1
Did Not Respond	1

GRADES 6 AND 7 READING

Current Position and Years of Experience in Education

		Years of Professional Experience in Education						
		Omit	1–5 years	6–10 years	11–15 years	16–20 years	More Than 20 years	Total
Current Position	Administrator	0	0	0	1	2	0	**3**
	Teacher	0	1	1	2	4	4	**12**
	Other	0	0	0	0	0	1	**1**
	Total	**0**	**1**	**1**	**3**	**6**	**5**	**16**

Gender Distribution

Gender	N-Count
Female	12
Male	4

Ethnicity Distribution

Ethnicity	N-Count
African American	1
Asian or Pacific Islander	0
Hispanic	1
Multi-racial	4
Native American	0
White	10

Experience with Student Populations

Student Population	N-Count
General Education	16
Special Education	14
English Language Learners	14
Low Socioeconomic Status	15

District Type

Type	N-Count
Metro	3
Suburban	7
Rural	5
Other	0
Did Not Respond	1

District Size

Type	N-Count
Large	7
Medium	5
Small	3
Other	0
Did Not Respond	1

District Socioeconomic Status

Type	N-Count
High	2
Moderate	7
Low	5
Other	1
Did Not Respond	1

GRADE 5 MATHEMATICS

Current Position and Years of Experience in Education

		Years of Professional Experience in Education						
		Omit	1–5 years	6–10 years	11–15 years	16–20 years	More Than 20 years	Total
Current Position	Administrator	0	0	1	0	0	1	**2**
	Teacher	0	1	1	2	1	2	**7**
	Other	0	0	0	1	1	1	**3**
	Total	**0**	**1**	**2**	**3**	**2**	**4**	**12**

Gender Distribution

Gender	N-Count
Female	8
Male	4

Ethnicity Distribution

Ethnicity	N-Count
African American	2
Asian or Pacific Islander	1
Hispanic	0
Multi-racial	5
Native American	0
White	4

Experience with Student Populations

Student Population	N-Count
General Education	12
Special Education	6
English Language Learners	11
Low Socioeconomic Status	11

District Type

Type	N-Count
Metro	2
Suburban	3
Rural	5
Other	1
Did Not Respond	1

District Size

Type	N-Count
Large	3
Medium	7
Small	1
Other	0
Did Not Respond	1

District Socioeconomic Status

Type	N-Count
High	0
Moderate	3
Low	7
Other	1
Did Not Respond	1

GRADE 5 ENGLISH READING

Current Position and Years of Experience in Education

		Years of Professional Experience in Education						
		Omit	1–5 years	6–10 years	11–15 years	16–20 years	More Than 20 years	Total
Current Position	Administrator	0	0	0	2	0	0	**2**
	Teacher	0	0	4	4	1	2	**11**
	Other	0	0	0	0	0	1	**1**
	Total	**0**	**0**	**4**	**6**	**1**	**3**	**14**

Gender Distribution

Gender	N-Count
Female	11
Male	3

Ethnicity Distribution

Ethnicity	N-Count
African American	3
Asian or Pacific Islander	1
Hispanic	1
Multi-racial	2
Native American	0
White	7

Experience with Student Populations

Student Population	N-Count
General Education	14
Special Education	12
English Language Learners	13
Low Socioeconomic Status	14

District Type

Type	N-Count
Metro	1
Suburban	9
Rural	4
Other	0
Did Not Respond	0

District Size

Type	N-Count
Large	6
Medium	7
Small	1
Other	0
Did Not Respond	0

District Socioeconomic Status

Type	N-Count
High	0
Moderate	9
Low	5
Other	0
Did Not Respond	0

GRADE 5 SPANISH READING

Current Position and Years of Experience in Education

		Years of Professional Experience in Education						
		Omit	1–5 years	6–10 years	11–15 years	16–20 years	More Than 20 years	Total
Current Position	Administrator	0	0	0	0	1	0	**1**
	Teacher	0	0	1	1	2	0	**4**
	Other	0	0	0	4	1	2	**7**
	Total	**0**	**0**	**1**	**5**	**4**	**2**	**12**

Gender Distribution

Gender	N-Count
Female	11
Male	1

Ethnicity Distribution

Ethnicity	N-Count
African American	0
Asian or Pacific Islander	0
Hispanic	2
Multi-racial	10
Native American	0
White	0

Experience with Student Populations

Student Population	N-Count
General Education	12
Special Education	12
English Language Learners	12
Low Socioeconomic Status	12

District Type

Type	N-Count
Metro	6
Suburban	5
Rural	1
Other	0
Did Not Respond	0

District Size

Type	N-Count
Large	9
Medium	1
Small	2
Other	0
Did Not Respond	0

District Socioeconomic Status

Type	N-Count
High	2
Moderate	2
Low	6
Other	2
Did Not Respond	0

GRADE 5 SCIENCE

Current Position and Years of Experience in Education

		Years of Professional Experience in Education						
		Omit	1–5 years	6–10 years	11–15 years	16–20 years	More Than 20 years	Total
Current Position	Administrator	0	0	0	0	0	0	**0**
	Teacher	0	0	4	2	1	4	**11**
	Other	0	0	0	0	0	0	**0**
	Total	**0**	**0**	**4**	**2**	**1**	**4**	**11**

Gender Distribution

Gender	N-Count
Female	10
Male	1

Ethnicity Distribution

Ethnicity	N-Count
African American	1
Asian or Pacific Islander	0
Hispanic	1
Multi-racial	3
Native American	0
White	6

Experience with Student Populations

Student Population	N-Count
General Education	11
Special Education	11
English Language Learners	10
Low Socioeconomic Status	11

District Type

Type	N-Count
Metro	2
Suburban	4
Rural	5
Other	0
Did Not Respond	0

District Size

Type	N-Count
Large	3
Medium	4
Small	4
Other	0
Did Not Respond	0

District Socioeconomic Status

Type	N-Count
High	1
Moderate	1
Low	8
Other	1
Did Not Respond	0

GRADE 4 ENGLISH WRITING (OCTOBER 2012)

Current Position and Years of Experience in Education

		Years of Professional Experience in Education						
		Omit	1–5 years	6–10 years	11–15 years	16–20 years	More Than 20 years	Total
Current Position	Administrator	0	0	0	0	0	1	**1**
	Teacher	0	1	0	1	0	3	**5**
	Other	0	0	0	2	2	1	**5**
	Total	**0**	**1**	**0**	**3**	**2**	**5**	**11**

Gender Distribution

Gender	N-Count
Female	9
Male	2

Ethnicity Distribution

Ethnicity	N-Count
African American	2
Asian or Pacific Islander	0
Hispanic	0
Multi-racial	2
Native American	0
White	7

Experience with Student Populations

Student Population	N-Count
General Education	11
Special Education	10
English Language Learners	11
Low Socioeconomic Status	10

District Type

Type	N-Count
Metro	1
Suburban	5
Rural	4
Other	0
Did Not Respond	1

District Size

Type	N-Count
Large	5
Medium	5
Small	0
Other	0
Did Not Respond	1

District Socioeconomic Status

Type	N-Count
High	3
Moderate	2
Low	5
Other	0
Did Not Respond	1

GRADE 4 ENGLISH WRITING (NOVEMBER 2012)

Current Position and Years of Experience in Education

		Years of Professional Experience in Education						
		Omit	1–5 years	6–10 years	11–15 years	16–20 years	More Than 20 years	Total
Current Position	Administrator	0	0	0	0	1	0	**1**
	Teacher	0	1	0	3	1	3	**8**
	Other	0	0	0	0	0	1	**1**
	Total	**0**	**1**	**0**	**3**	**2**	**4**	**10**

Gender Distribution

Gender	N-Count
Female	8
Male	2

Ethnicity Distribution

Ethnicity	N-Count
African American	2
Asian or Pacific Islander	0
Hispanic	0
Multi-racial	2
Native American	0
White	6

Experience with Student Populations

Student Population	N-Count
General Education	10
Special Education	10
English Language Learners	10
Low Socioeconomic Status	9

District Type

Type	N-Count
Metro	1
Suburban	5
Rural	3
Other	0
Did Not Respond	1

District Size

Type	N-Count
Large	5
Medium	4
Small	0
Other	0
Did Not Respond	1

District Socioeconomic Status

Type	N-Count
High	2
Moderate	2
Low	5
Other	0
Did Not Respond	1

GRADE 4 SPANISH WRITING (OCTOBER 2012)

Current Position and Years of Experience in Education

		Years of Professional Experience in Education						
		Omit	1–5 years	6–10 years	11–15 years	16–20 years	More Than 20 years	Total
Current Position	Administrator	0	0	0	0	0	0	**0**
	Teacher	0	1	4	4	3	0	**12**
	Other	0	0	1	0	0	0	**1**
	Total	**0**	**1**	**5**	**4**	**3**	**0**	**13**

Gender Distribution

Gender	N-Count
Female	10
Male	3

Ethnicity Distribution

Ethnicity	N-Count
African American	0
Asian or Pacific Islander	0
Hispanic	4
Multi-racial	7
Native American	0
White	2

Experience with Student Populations

Student Population	N-Count
General Education	11
Special Education	9
English Language Learners	12
Low Socioeconomic Status	12

District Type

Type	N-Count
Metro	4
Suburban	5
Rural	3
Other	1
Did Not Respond	0

District Size

Type	N-Count
Large	2
Medium	9
Small	2
Other	0
Did Not Respond	0

District Socioeconomic Status

Type	N-Count
High	0
Moderate	8
Low	5
Other	0
Did Not Respond	0

GRADE 4 SPANISH WRITING (NOVEMBER 2012)

Current Position and Years of Experience in Education

		Years of Professional Experience in Education						
		Omit	1–5 years	6–10 years	11–15 years	16–20 years	More Than 20 years	Total
Current Position	Administrator	0	0	0	0	0	0	**0**
	Teacher	0	0	5	2	3	0	**10**
	Other	0	0	0	0	0	0	**0**
	Total	**0**	**0**	**5**	**2**	**3**	**0**	**10**

Gender Distribution

Gender	N-Count
Female	8
Male	2

Ethnicity Distribution

Ethnicity	N-Count
African American	0
Asian or Pacific Islander	0
Hispanic	4
Multi-racial	5
Native American	0
White	1

Experience with Student Populations

Student Population	N-Count
General Education	8
Special Education	6
English Language Learners	9
Low Socioeconomic Status	9

District Type

Type	N-Count
Metro	3
Suburban	4
Rural	2
Other	1
Did Not Respond	0

District Size

Type	N-Count
Large	2
Medium	7
Small	1
Other	0
Did Not Respond	0

District Socioeconomic Status

Type	N-Count
High	0
Moderate	7
Low	3
Other	0
Did Not Respond	0

GRADES 3 AND 4 MATHEMATICS

Current Position and Years of Experience in Education

		Years of Professional Experience in Education						
		Omit	1–5 years	6–10 years	11–15 years	16–20 years	More Than 20 years	Total
Current Position	Administrator	0	0	0	0	0	0	**0**
	Teacher	0	1	1	2	5	5	**14**
	Other	0	0	0	0	1	0	**1**
	Total	**0**	**1**	**1**	**2**	**6**	**5**	**15**

Gender Distribution

Gender	N-Count
Female	13
Male	2

Ethnicity Distribution

Ethnicity	N-Count
African American	2
Asian or Pacific Islander	0
Hispanic	0
Multi-racial	4
Native American	0
White	9

Experience with Student Populations

Student Population	N-Count
General Education	14
Special Education	14
English Language Learners	11
Low Socioeconomic Status	15

District Type

Type	N-Count
Metro	2
Suburban	3
Rural	7
Other	1
Did Not Respond	2

District Size

Type	N-Count
Large	3
Medium	5
Small	5
Other	0
Did Not Respond	2

District Socioeconomic Status

Type	N-Count
High	0
Moderate	3
Low	9
Other	1
Did Not Respond	2

GRADES 3 AND 4 ENGLISH READING

Current Position and Years of Experience in Education

		Years of Professional Experience in Education						
		Omit	1–5 years	6–10 years	11–15 years	16–20 years	More Than 20 years	Total
Current Position	Administrator	0	0	0	1	0	0	**1**
	Teacher	0	0	5	4	1	2	**12**
	Other	0	0	0	0	1	2	**3**
	Total	**0**	**0**	**5**	**5**	**2**	**4**	**16**

Gender Distribution

Gender	N-Count
Female	14
Male	2

Ethnicity Distribution

Ethnicity	N-Count
African American	4
Asian or Pacific Islander	0
Hispanic	0
Multi-racial	2
Native American	0
White	9
Did Not Respond	1

Experience with Student Populations

Student Population	N-Count
General Education	16
Special Education	14
English Language Learners	13
Low Socioeconomic Status	16

District Type

Type	N-Count
Metro	1
Suburban	6
Rural	8
Other	0
Did Not Respond	1

District Size

Type	N-Count
Large	3
Medium	8
Small	5
Other	0
Did Not Respond	0

District Socioeconomic Status

Type	N-Count
High	1
Moderate	6
Low	9
Other	0
Did Not Respond	0

GRADES 3 AND 4 SPANISH READING

Current Position and Years of Experience in Education

		Years of Professional Experience in Education						
		Omit	1–5 years	6–10 years	11–15 years	16–20 years	More Than 20 years	Total
Current Position	Administrator	0	0	0	0	1	0	**1**
	Teacher	0	2	3	5	0	2	**12**
	Other	0	0	0	1	1	0	**2**
	Total	**0**	**2**	**3**	**6**	**2**	**2**	**15**

Gender Distribution

Gender	N-Count
Female	9
Male	6

Ethnicity Distribution

Ethnicity	N-Count
African American	0
Asian or Pacific Islander	0
Hispanic	3
Multi-racial	11
Native American	0
White	1

Experience with Student Populations

Student Population	N-Count
General Education	12
Special Education	11
English Language Learners	13
Low Socioeconomic Status	14

District Type

Type	N-Count
Metro	5
Suburban	6
Rural	4
Other	0
Did Not Respond	0

District Size

Type	N-Count
Large	7
Medium	6
Small	2
Other	0
Did Not Respond	0

District Socioeconomic Status

Type	N-Count
High	0
Moderate	7
Low	8
Other	0
Did Not Respond	0

Appendix 14: Example Standard Setting Feedback Data

This appendix provides examples of the committee-level feedback data that were presented to the standard-setting panelists after each round of judgment. The examples given are for the STAAR English I reading assessment and the STAAR grade 5 mathematics assessment. Similar types of feedback data were provided for the other STAAR assessments.

For a complete summary of the panelist judgments and standard-setting meeting outcomes for each STAAR assessment, refer to Appendices 16–19.

STAAR EOC – English I Reading

Round 1 Feedback Data

STAAR English I Reading—Round 1

Performance Standard	Level II	Level III
Minimum Page Number	39	50
Maximum Page Number	46	57
Mean Page Number	41.8	54.2
Median Page Number	42	54

Figure A14.1: Summary of Cut Score Recommendations (Bookmarked Page Numbers)

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 A14.2: Likelihood Table Based on Cut Score Recommendations (Median Page Numbers)

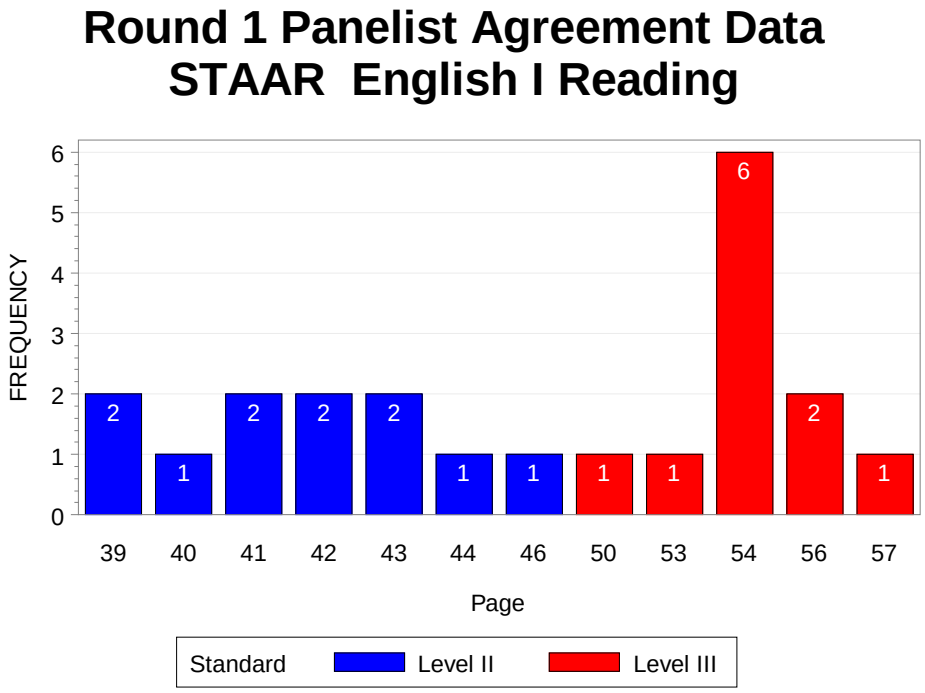


Figure A14.3: Cut Score Recommendation (Bookmarked Page Numbers) Distribution

ROUNDS 2 AND 3 FEEDBACK DATA[3]

STAAR English I Reading—Round 2

Performance Standard	Level II	Level III
Minimum Page Number	37	41
Maximum Page Number	44	56
Mean Page Number	41.3	52.6
Median Page Number	41	54

Figure A14.4: Summary of Cut Score Recommendations (Bookmarked Page Numbers)

[3] The same types of feedback data were generated after Rounds 2 and 3. Examples for Round 2 feedback data are provided.

STAAR English I Reading—Round 2

Performance Standard	Level II	Level III
Borderline Student		
Probability of reaching the corresponding cut in English II Reading	50%	24%
Typical Student		
Probability of reaching the corresponding cut in English II Reading	72%	49%

Figure A14.5: Likelihood Table Based on Cut Score Recommendations (Median Page Numbers)

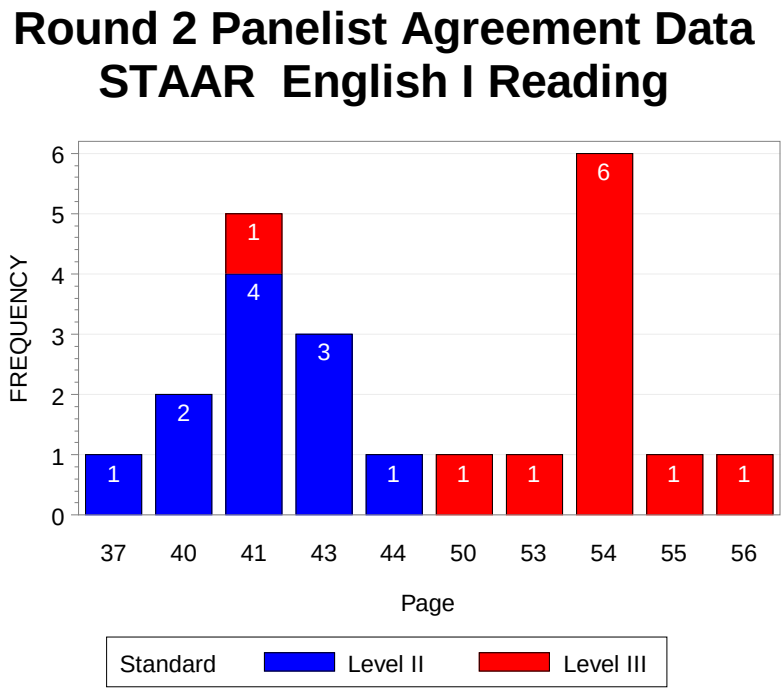
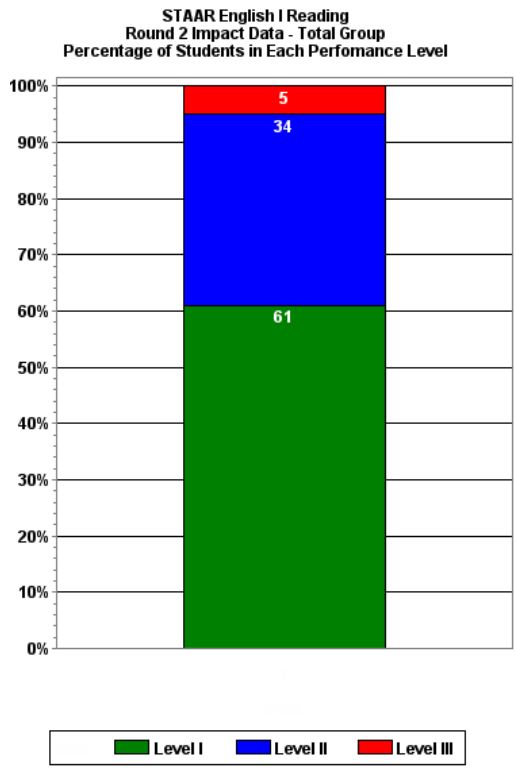


Figure A14.6: Cut Score Recommendation (Bookmarked Page Numbers) Distribution



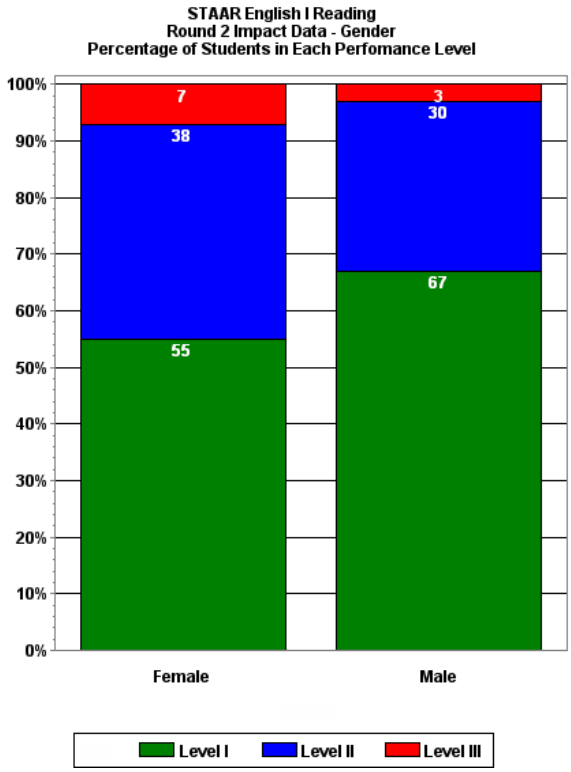


Figure A14.7: Impact Data (Total Group and By Gender) Based on Cut Score Recommendations

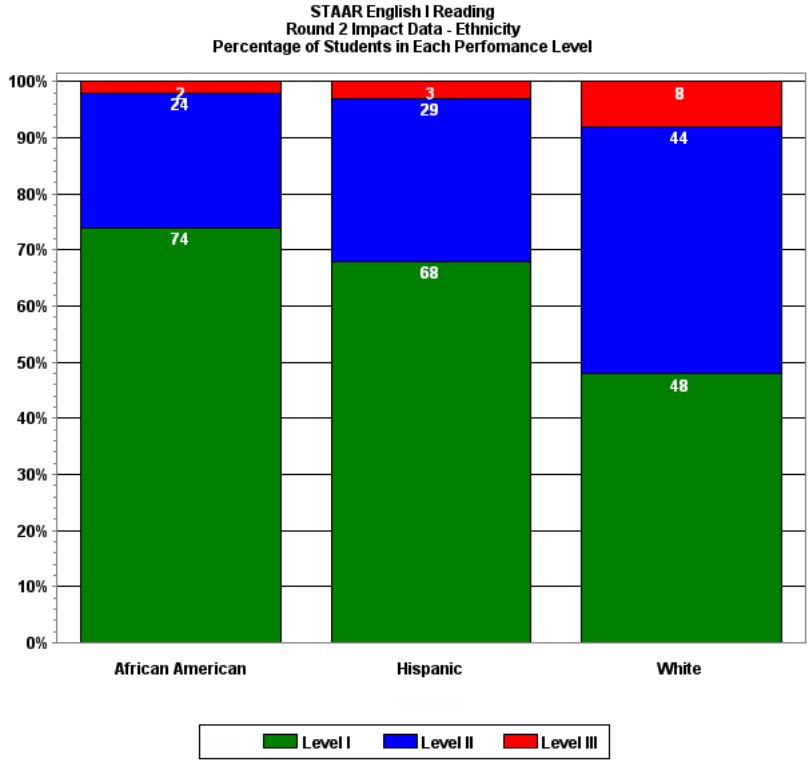


Figure A14.8: Impact Data (By Ethnicity) Based on Cut Score Recommendations

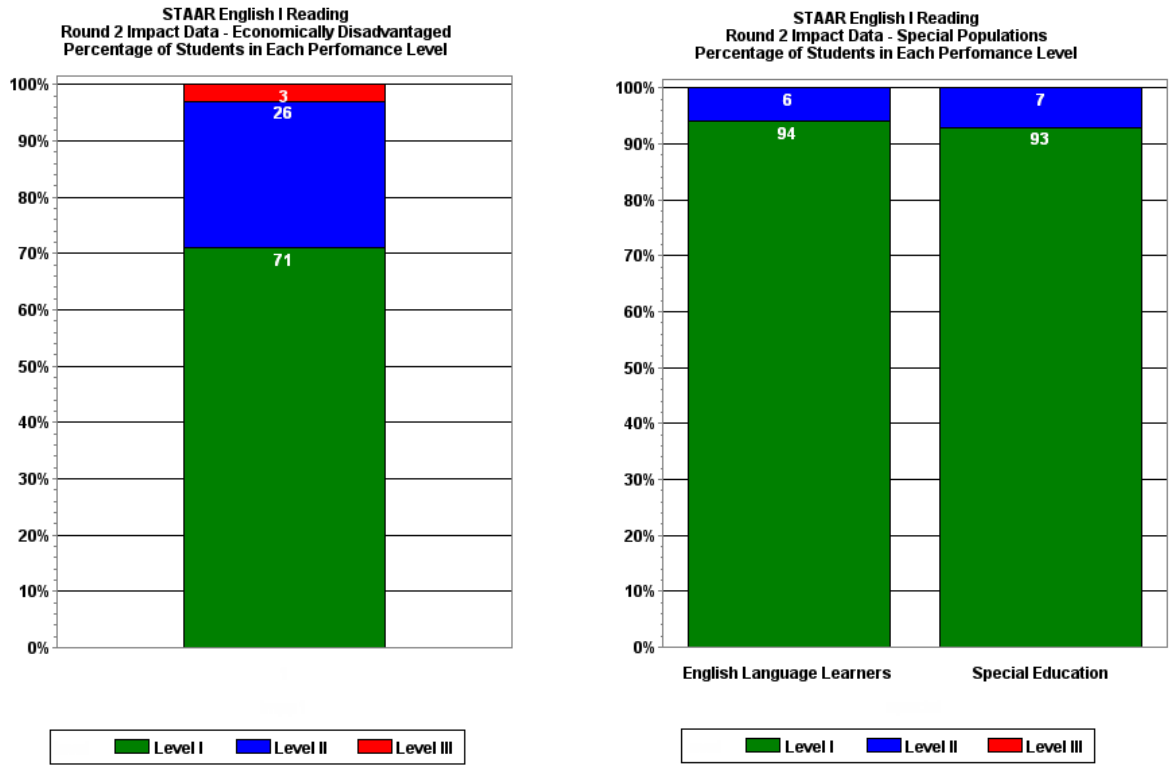


Figure A14.9: Impact Data (Economically Disadvantaged and Special Populations) Based on Cut Score Recommendations

STAAR 3–8 – Grade 5 Mathematics

ROUND 1 FEEDBACK DATA

STAAR Grade 5 Mathematics—Round 1

Performance Standard	Level II	Level III
Minimum Page Number	38	56
Maximum Page Number	54	68
Mean Page Number	43.3	61.2
Median Page Number	41	60

Figure A14.10: Summary of Cut Score Recommendations (Bookmarked Page Numbers)

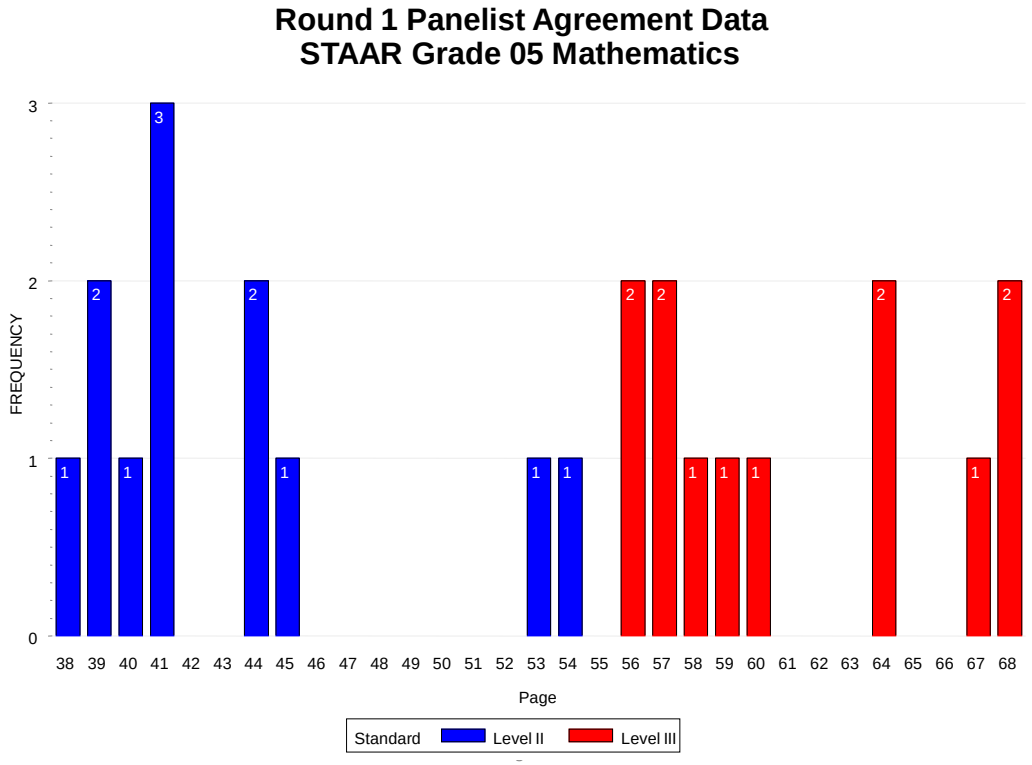


Figure A14.11: Cut Score Recommendation (Bookmarked Page Numbers) Distribution

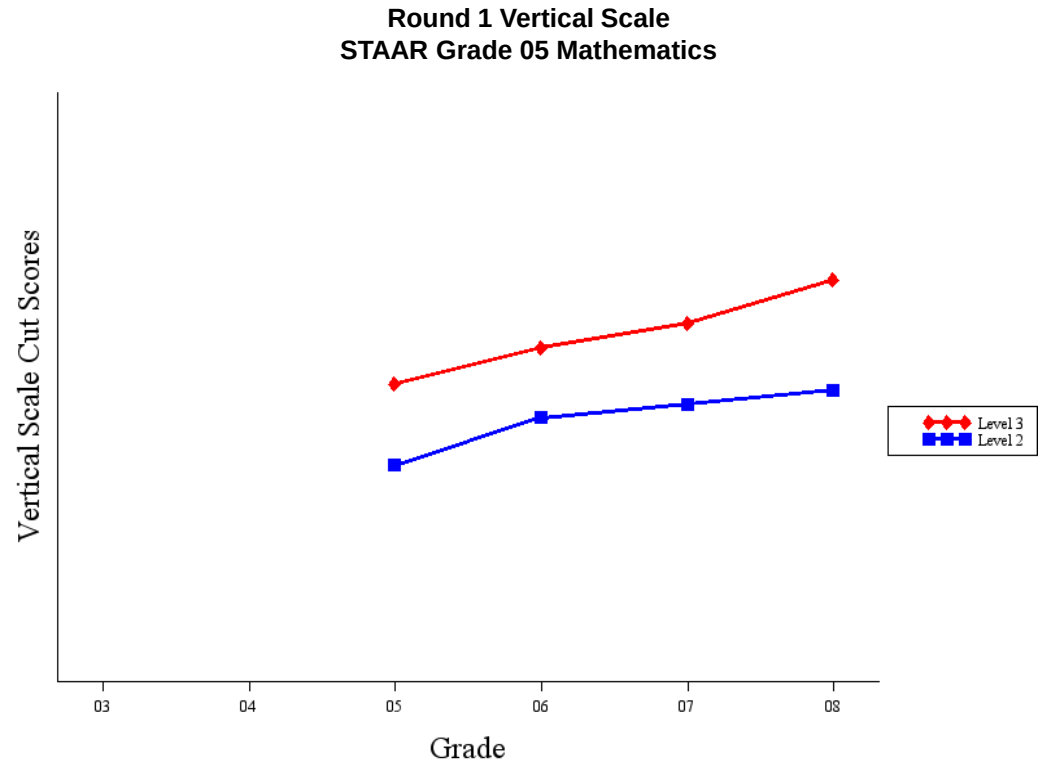


Figure A14.12: Cut Score Recommendation for Grades 5–8 (Vertical Scale)

STAAR Grade 5 Mathematics—Round 2

Performance Standard	Level II	Level III
Minimum Page Number	39	56
Maximum Page Number	53	65
Mean Page Number	44.2	61.1
Median Page Number	43	63

Figure A14.13: Summary of Cut Score Recommendations (Bookmarked Page Numbers)

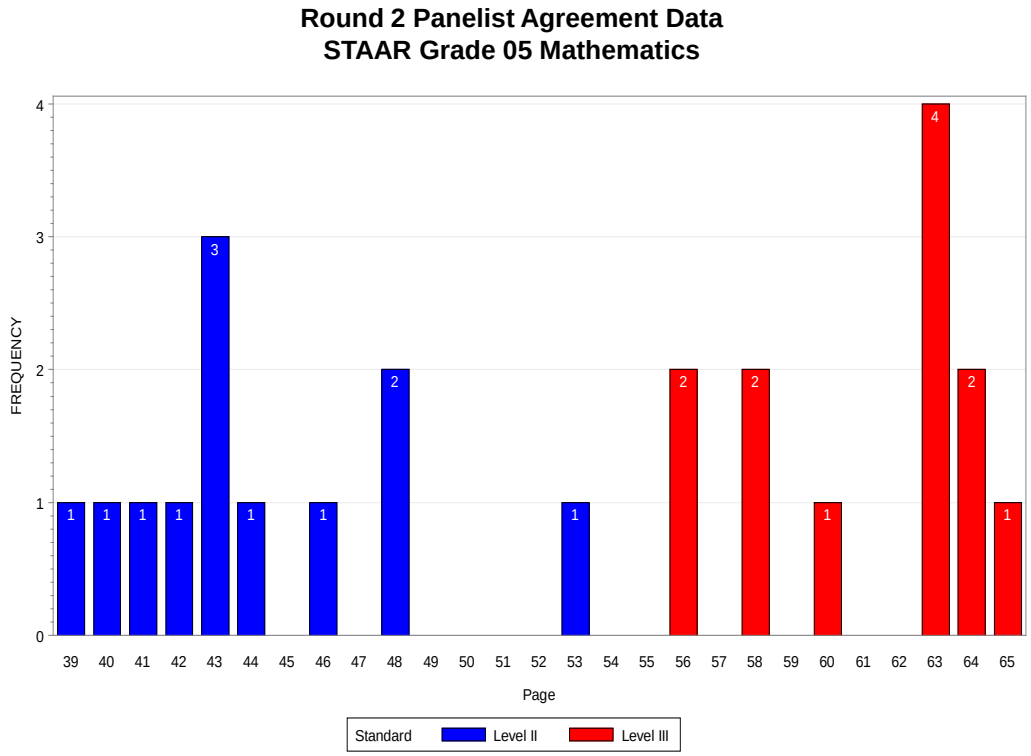


Figure A14.14: Cut Score Recommendation (Bookmarked Page Numbers) Distribution

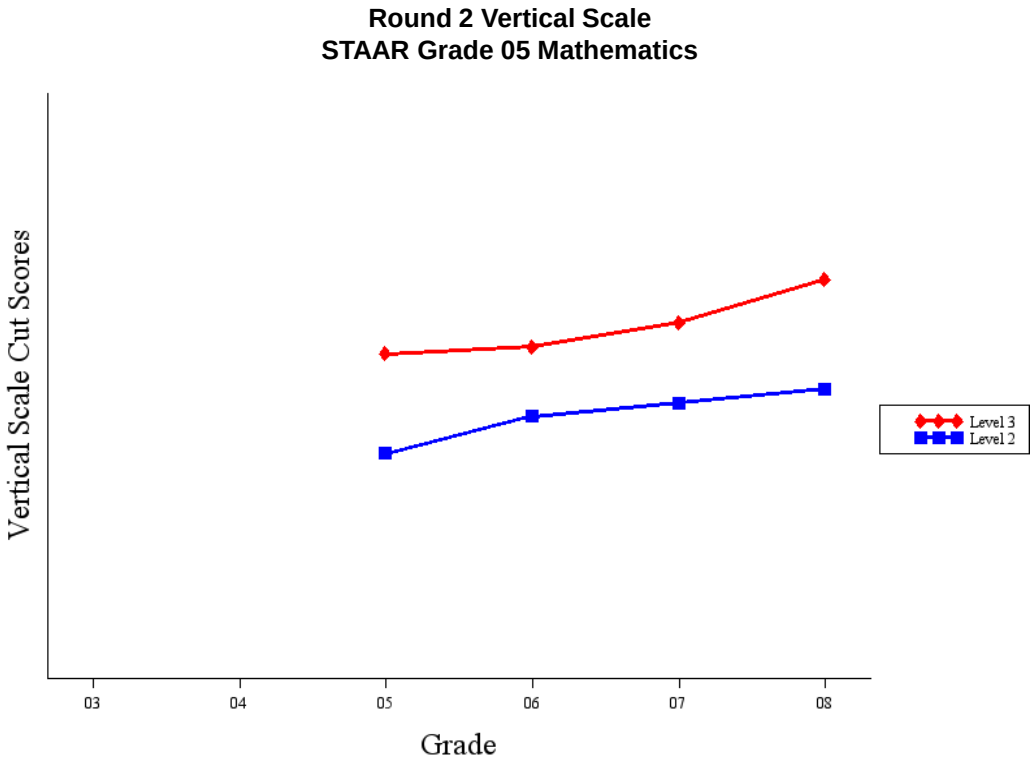


Figure A14.15: Cut Score Recommendation for Grades 5–8 (Vertical Scale)

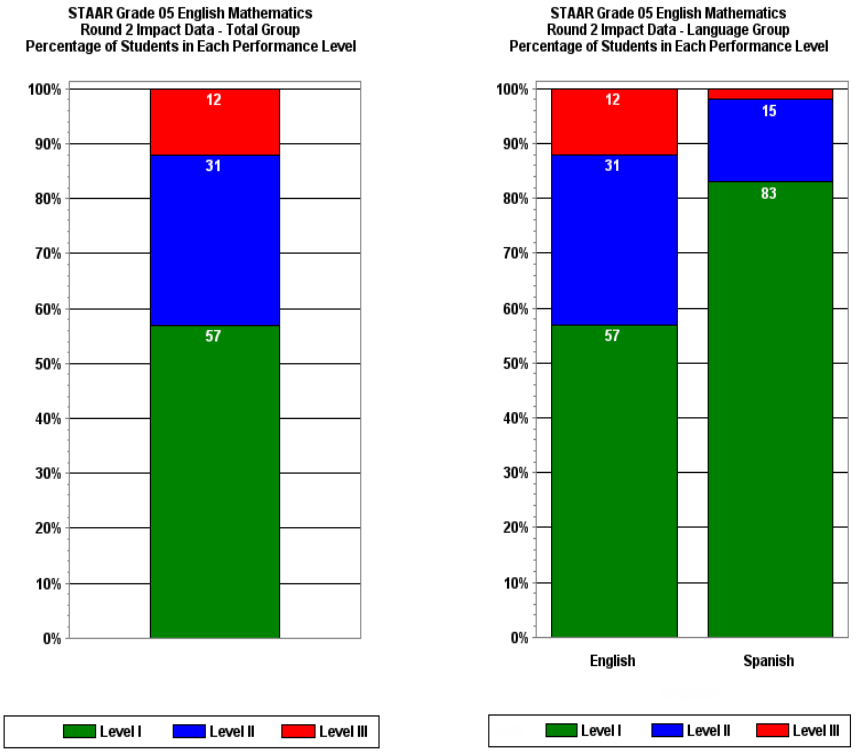


Figure A14.16: Impact Data (Total Group and By Gender) Based on Cut Score Recommendations

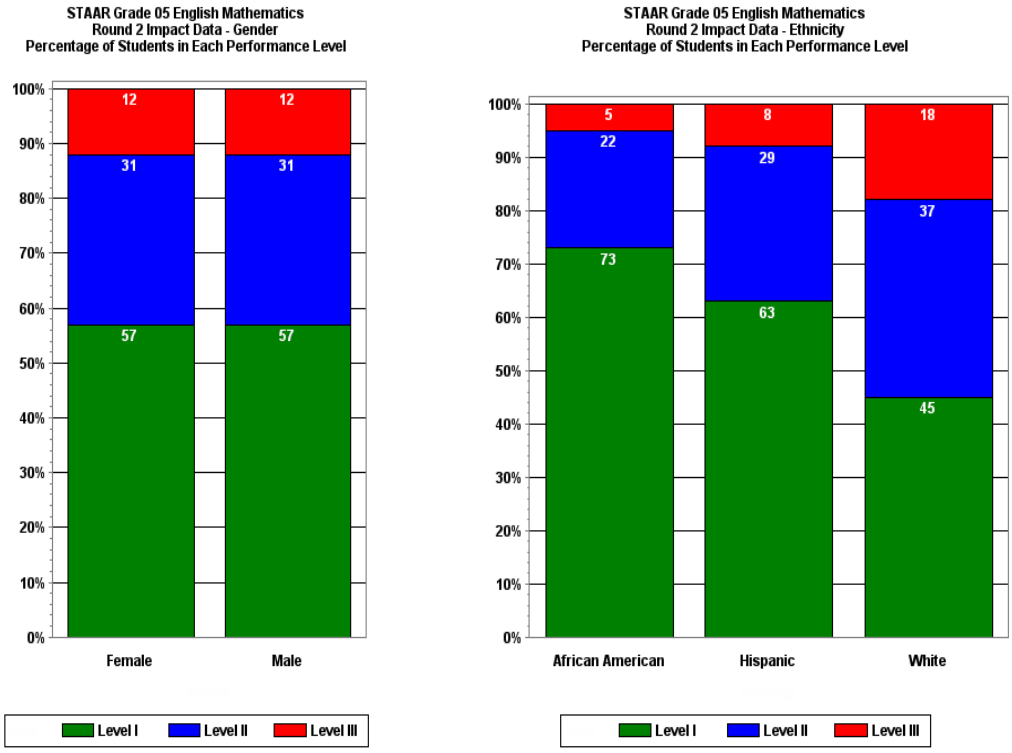


Figure A14.17: Impact Data (By Ethnicity) Based on Cut Score Recommendations

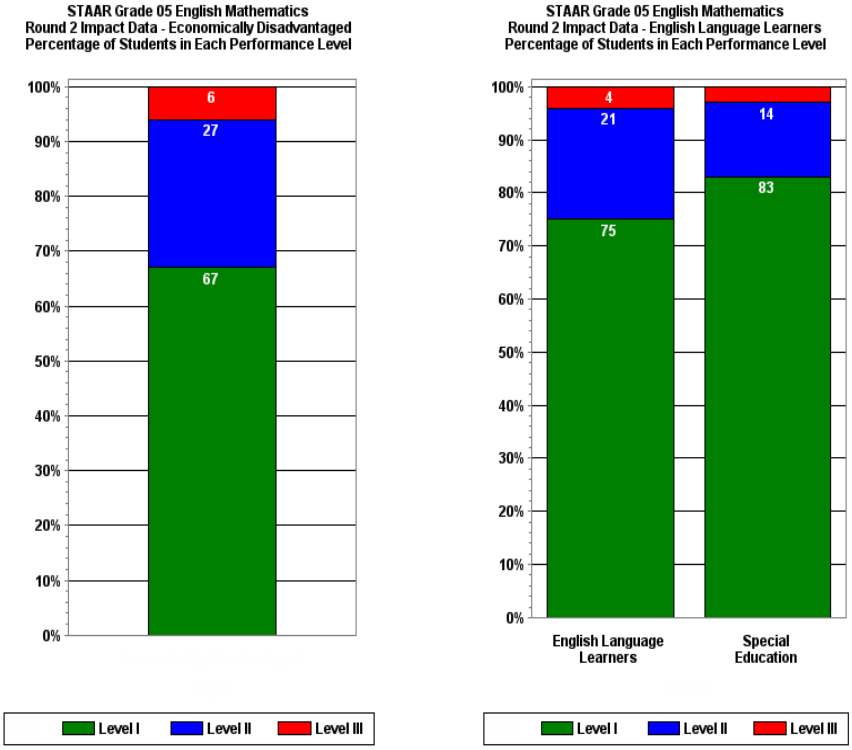


Figure A14.18: Impact Data (Economically Disadvantaged and Special Populations) Based on Cut Score Recommendations

STAAR Grade 5 Mathematics—Round 3

Performance Standard	Level II	Level III
Minimum Page Number	41	58
Maximum Page Number	54	65
Mean Page Number	47.4	62.0
Median Page Number	48	63

Figure A14.19: Summary of Cut Score Recommendations (Bookmarked Page Numbers)

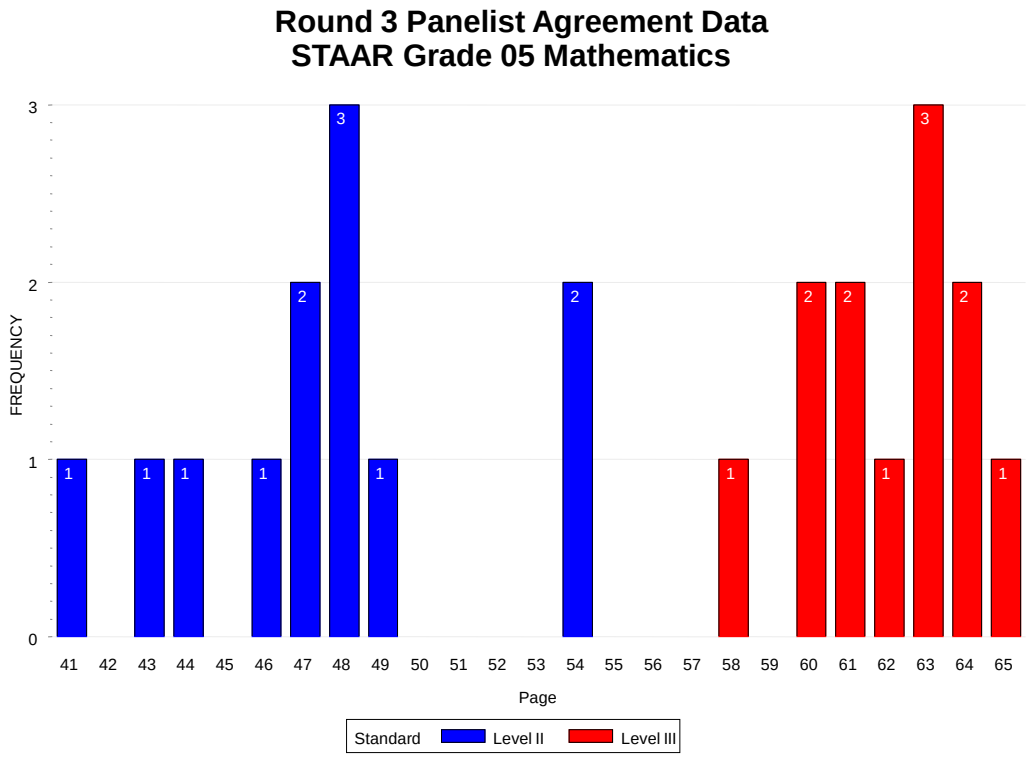


Figure A14.20: Cut Score Recommendation (Bookmarked Page Numbers) Distribution

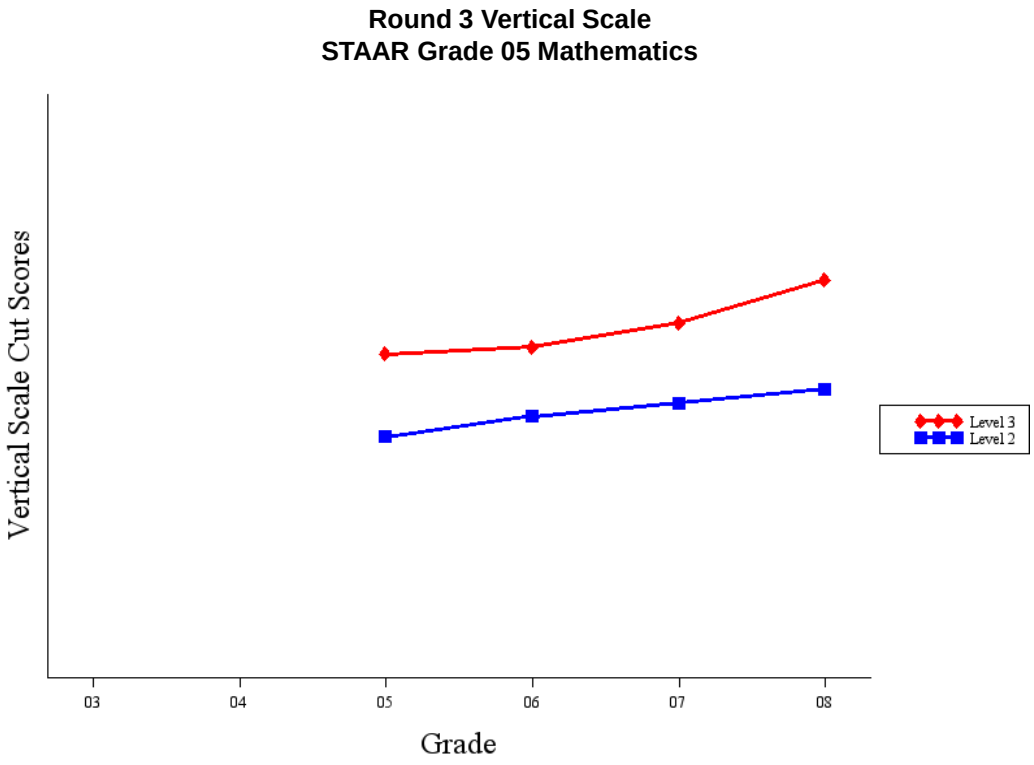


Figure A14.21: Cut Score Recommendation (Bookmarked Page Numbers) Distribution

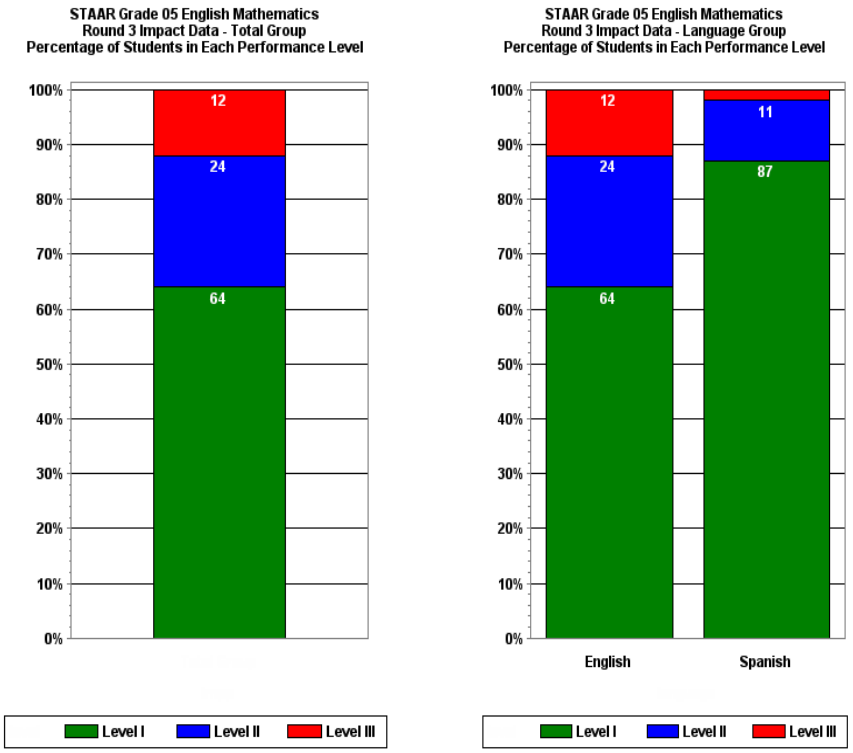
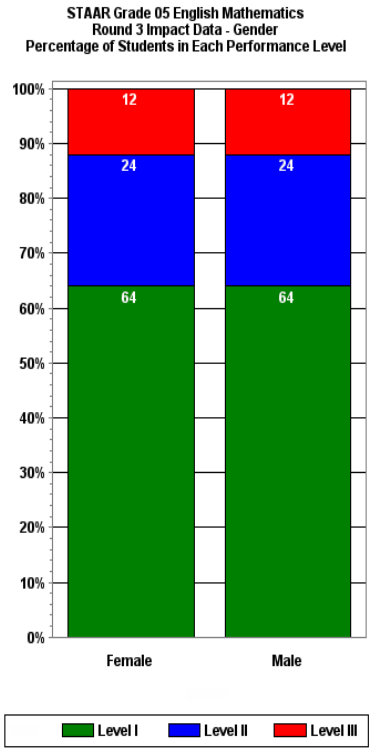


Figure A14.22: Impact Data (Total Group and By Gender) Based on Cut Score Recommendations



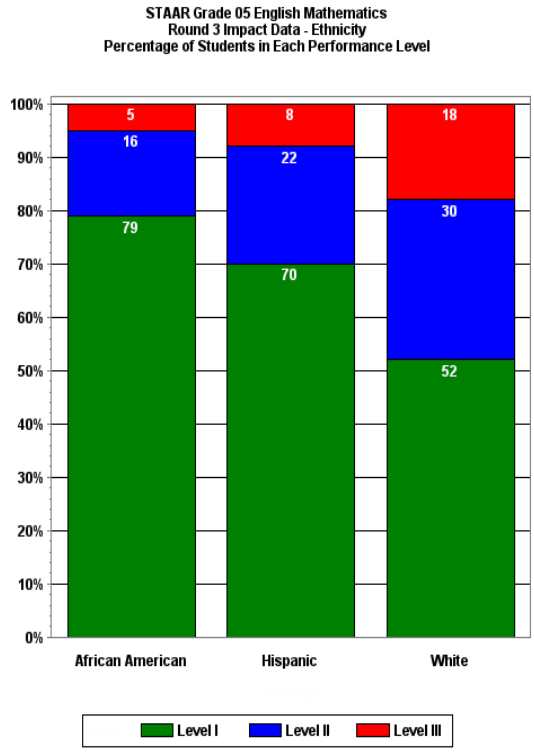
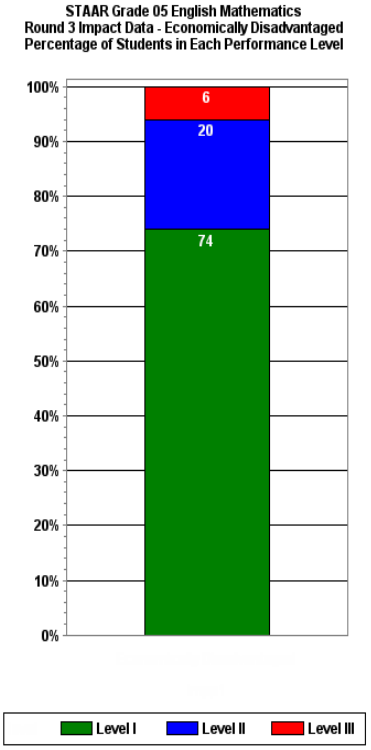


Figure A14.23: Impact Data (By Ethnicity) Based on Cut Score Recommendations



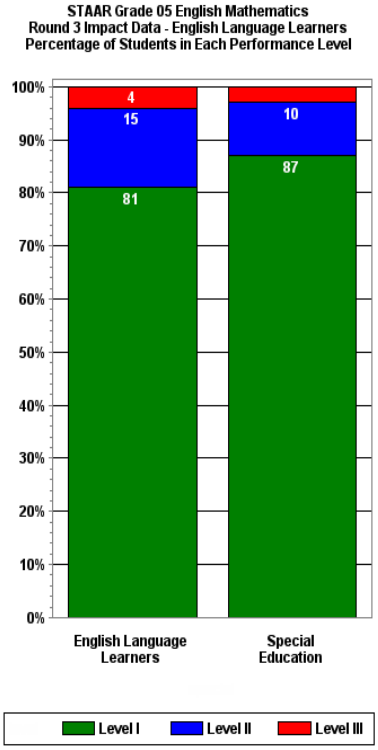


Figure A14.24: Impact Data (Economically Disadvantaged and Special Populations) Based on Cut Score Recommendations

Appendix 15: Standard-Setting Process Evaluation Summary

SECTION INSTRUCTIONS

The instructions provided for each section of the process evaluation survey for the standard-setting committee meetings are as follows.

Section 1 (Meeting Success): Check the column below that best reflects your opinion about the level of success of the various components of the meeting in which you have just participated. The activities were designed to help you both understand the process and be supportive of the recommendations made by the committee.
Section 2 (Usefulness of Activities and Information): How useful do you feel the following activities and/or information were in assisting you to make your recommendations?
Section 3 (Adequacy of Meeting Elements): How adequate were the following elements of the session?
Section 4 (Specific PLDs): In applying the standard-setting method, you were asked to recommend cut scores (separating three performance levels) for student performance on STAAR assessments. How confident do you feel that the specific Performance Level Descriptors (PLDs) are reasonable for each student performance level?
Section 5 (Cut Score Recommendations): How confident do you feel that the final cut score recommendations represent appropriate levels of student performance?
Section 6 (Opportunities to Express Opinions): Did you have adequate opportunities during the session to do the following?
Section 7 (Respect): Do you believe your opinions and judgments were treated with respect by the following?

A summary of responses given by each standard-setting committee is provided in the sections below. Please click on the link to go directly to the section of interest.

STAAR 3–8 Committees
Grade 8 Mathematics
Grade 8 Reading
Grade 8 Science
Grade 8 Social Studies
Grade 7 Writing (October 2012)
Grade 7 Writing (November 2012)
Grades 6 and 7 Mathematics
Grades 6 and 7 Reading
Grade 5 Mathematics
Grade 5 English Reading
Grade 5 Spanish Reading
Grade 5 Science
Grade 4 English Writing (October 2012)
Grade 4 English Writing (November 2012)
Grade 4 Spanish Writing (October 2012)
Grade 4 Spanish Writing (November 2012)
Grades 3 and 4 Mathematics
Grades 3 and 4 English Reading
Grades 3 and 4 Spanish Reading

ENGLISH I, II, AND III READING

A total of 10 panelists responded to the process evaluation survey.

Section 1: Meeting Success

Meeting Component	Not Successful	Partially Successful	Successful	Very Successful	Omit
Introduction to the process of setting performance standards	0%	0%	20%	80%	0%
Discussion of the performance labels and the definitions	0%	0%	0%	100%	0%
Taking the actual assessment(s)	0%	0%	0%	100%	0%
Overview of the item mapping procedure	0%	0%	10%	90%	0%
Practice exercise for the item-mapping procedure	0%	10%	20%	70%	0%
Feedback data provided in each round	0%	0%	20%	80%	0%
Discussion after each round	0%	10%	20%	70%	0%
Articulation	0%	10%	40%	40%	10%

Section 2: Usefulness of Activities and Information

Activity or Information	Not Useful	Somewhat Useful	Useful	Very Useful	Omit
Specific Performance Level Descriptors (PLDs)	0%	0%	30%	70%	0%
Training in the bookmark standard setting method	0%	0%	0%	100%	0%
Feedback data provided after Round 1	0%	0%	10%	90%	0%
Feedback data provided after Round 2	0%	0%	30%	70%	0%
Presentation of data across courses	0%	0%	10%	90%	0%

Section 3: Adequacy of Meeting Elements

Meeting Element	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Training provided	0%	0%	0%	100%	0%
Amount of time spent training	0%	0%	30%	70%	0%
Feedback provided between rounds	0%	0%	20%	80%	0%
Facilities used for the session	0%	0%	10%	90%	0%
Total amount of time in breakout groups to make judgments	0%	0%	20%	80%	0%
Number of rounds for the judgments	0%	0%	40%	50%	10%

Section 4: Specific PLDs

Performance Category	Not Confident	Somewhat Confident	Confident	Very Confident	Omit
Level I: Unsatisfactory Academic Performance	0%	0%	50%	40%	10%
Level II: Satisfactory Academic Performance	0%	0%	50%	40%	10%
Level III: Advanced Academic Performance	0%	0%	60%	30%	10%

Section 5: Cut Score Recommendations

Cut Score	Not Confident	Somewhat Confident	Confident	Very Confident	Omit
Level II: Satisfactory Academic Performance	0%	10%	40%	40%	10%
Level III: Advanced Academic Performance	0%	10%	50%	30%	10%

Section 6: Opportunities to Express Opinions

Category	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Express your opinions about student performance levels	0%	0%	10%	80%	10%
Ask questions about the standards and how they will be used	0%	10%	20%	60%	10%
Ask questions about the process of making cut score recommendations	0%	0%	20%	70%	10%
Interact with your fellow panelists	0%	0%	20%	80%	0%

Section 7: Respect

Party	No	Sometimes	Yes	Omit
Fellow panelists	0%	0%	100%	0%
Facilitators	0%	0%	100%	0%

ENGLISH I, II, AND III WRITING

A total of 15 panelists responded to the process evaluation survey.

Section 1: Meeting Success

Meeting Component	Not Successful	Partially Successful	Successful	Very Successful	Omit
Introduction to the process of setting performance standards	0%	0%	27%	73%	0%
Discussion of the performance labels and the definitions	0%	0%	20%	80%	0%
Taking the actual assessment(s)	0%	0%	20%	80%	0%
Overview of the item mapping procedure	0%	0%	20%	80%	0%
Practice exercise for the item-mapping procedure	0%	0%	33%	67%	0%
Feedback data provided in each round	0%	0%	6%	93%	0%
Discussion after each round	0%	0%	20%	80%	0%
Articulation	0%	20%	53%	27%	0%

Section 2: Usefulness of Activities and Information

Activity or Information	Not Useful	Somewhat Useful	Useful	Very Useful	Omit
Specific Performance Level Descriptors (PLDs)	0%	13%	7%	80%	0%
Training in the bookmark standard setting method	0%	7%	13%	80%	0%
Feedback data provided after Round 1	0%	0%	13%	87%	0%
Feedback data provided after Round 2	0%	7%	0%	93%	0%
Presentation of data across courses	7%	13%	7%	73%	0%

Section 3: Adequacy of Meeting Elements

Meeting Element	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Training provided	0%	0%	33%	67%	0%
Amount of time spent training	0%	0%	40%	60%	0%
Feedback provided between rounds	0%	0%	20%	80%	0%
Facilities used for the session	0%	0%	7%	93%	0%
Total amount of time in breakout groups to make judgments	0%	0%	27%	73%	0%
Number of rounds for the judgments	0%	0%	20%	73%	7%

Section 4: Specific PLDs

Performance Category	Not Confident	Somewhat Confident	Confident	Very Confident	Omit
Level I: Unsatisfactory Academic Performance	0%	13%	7%	73%	7%
Level II: Satisfactory Academic Performance	0%	7%	20%	67%	6%
Level III: Advanced Academic Performance	0%	7%	20%	67%	6%

Section 5: Cut Score Recommendations

Cut Score	Not Confident	Somewhat Confident	Confident	Very Confident	Omit
Level II: Satisfactory Academic Performance	0%	7%	27%	60%	6%
Level III: Advanced Academic Performance	0%	7%	27%	60%	6%

Section 6: Opportunities to Express Opinions

Category	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Express your opinions about student performance levels	0%	0%	27%	67%	6%
Ask questions about the standards and how they will be used	0%	7%	33%	53%	6%
Ask questions about the process of making cut score recommendations	0%	0%	13%	80%	7%
Interact with your fellow panelists	0%	0%	20%	80%	0%

Section 7: Respect

Party	No	Sometimes	Yes	Omit
Fellow panelists	0%	13%	87%	0%
Facilitators	0%	7%	93%	0%

MATHEMATICS — ALGEBRA I, GEOMETRY, AND ALGEBRA II

A total of 26 panelists responded to the process evaluation survey.

Section 1: Meeting Success

Meeting Component	Not Successful	Partially Successful	Successful	Very Successful	Omit
Introduction to the process of setting performance standards	0%	0%	12%	88%	0%
Discussion of the performance labels and the definitions	0%	0%	12%	88%	0%
Taking the actual assessment(s)	0%	0%	4%	96%	0%
Overview of the item mapping procedure	0%	0%	11%	85%	4%
Practice exercise for the item-mapping procedure	0%	0%	8%	88%	4%
Feedback data provided in each round	0%	0%	8%	92%	0%
Discussion after each round	0%	0%	19%	77%	4%
Articulation	0%	4%	19%	73%	4%

Section 2: Usefulness of Activities and Information

Activity or Information	Not Useful	Somewhat Useful	Useful	Very Useful	Omit
Specific Performance Level Descriptors (PLDs)	0%	15%	20%	65%	0%
Training in the bookmark standard setting method	0%	8%	11%	81%	0%
Feedback data provided after Round 1	0%	0%	15%	85%	0%
Feedback data provided after Round 2	0%	0%	15%	85%	0%
Presentation of data across courses	0%	4%	15%	81%	0%

Section 3: Adequacy of Meeting Elements

Meeting Element	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Training provided	0%	0%	19%	81%	0%
Amount of time spent training	0%	0%	27%	73%	0%
Feedback provided between rounds	0%	0%	23%	77%	0%

Meeting Element	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Facilities used for the session	0%	0%	8%	92%	0%
Total amount of time in breakout groups to make judgments	0%	0%	19%	81%	0%
Number of rounds for the judgments	0%	0%	35%	65%	0%

Section 4: Specific PLDs

Performance Category	Not Confident	Somewhat Confident	Confident	Very Confident	Omit
Level I: Unsatisfactory Academic Performance	4%	8%	38%	50%	0%
Level II: Satisfactory Academic Performance	4%	12%	42%	38%	4%
Level III: Advanced Academic Performance	0%	4%	38%	58%	0%

Section 5: Cut Score Recommendations

Cut Score	Not Confident	Somewhat Confident	Confident	Very Confident	Omit
Level II: Satisfactory Academic Performance	4%	19%	42%	35%	0%
Level III: Advanced Academic Performance	4%	0%	42%	50%	4%

Section 6: Opportunities to Express Opinions

Category	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Express your opinions about student performance levels	0%	0%	23%	77%	0%
Ask questions about the standards and how they will be used	0%	0%	19%	77%	4%
Ask questions about the process of making cut score recommendations	0%	0%	19%	73%	8%
Interact with your fellow panelists	0%	0%	15%	85%	0%

Section 7: Respect

Party	No	Sometimes	Yes	Omit
Fellow panelists	0%	4%	96%	0%
Facilitators	0%	0%	100%	0%

SCIENCE — BIOLOGY

A total of 14 panelists responded to the process evaluation survey.

Section 1: Meeting Success

Meeting Component	Not Successful	Partially Successful	Successful	Very Successful	Omit
Introduction to the process of setting performance standards	7%	0%	36%	57%	0%
Discussion of the performance labels and the definitions	7%	0%	22%	71%	0%
Taking the actual assessment(s)	7%	0%	22%	71%	0%
Overview of the item mapping procedure	7%	0%	29%	64%	0%

Meeting Component	Not Successful	Partially Successful	Successful	Very Successful	Omit
Practice exercise for the item-mapping procedure	7%	0%	29%	57%	7%
Feedback data provided in each round	7%	0%	29%	64%	0%
Discussion after each round	7%	0%	29%	64%	0%
Articulation	7%	7%	29%	50%	7%

Section 2: Usefulness of Activities and Information

Activity or Information	Not Useful	Somewhat Useful	Useful	Very Useful	Omit
Specific Performance Level Descriptors (PLDs)	7%	0%	22%	71%	0%
Training in the bookmark standard setting method	7%	0%	43%	50%	0%
Feedback data provided after Round 1	7%	0%	22%	64%	7%
Feedback data provided after Round 2	7%	0%	22%	71%	0%
Presentation of data across courses	7%	0%	29%	64%	0%

Section 3: Adequacy of Meeting Elements

Meeting Element	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Training provided	7%	0%	36%	57%	0%
Amount of time spent training	7%	0%	36%	57%	0%
Feedback provided between rounds	7%	0%	36%	50%	7%
Facilities used for the session	7%	0%	7%	86%	0%
Total amount of time in breakout groups to make judgments	7%	0%	29%	64%	0%
Number of rounds for the judgments	7%	7%	71%	15%	0%

Section 4: Specific PLDs

Performance Category	Not Confident	Somewhat Confident	Confident	Very Confident	Omit
Level I: Unsatisfactory Academic Performance	7%	7%	50%	29%	7%
Level II: Satisfactory Academic Performance	7%	7%	57%	29%	0%
Level III: Advanced Academic Performance	7%	7%	50%	36%	0%

Section 5: Cut Score Recommendations

Cut Score	Not Confident	Somewhat Confident	Confident	Very Confident	Omit
Level II: Satisfactory Academic Performance	0%	36%	43%	21%	0%
Level III: Advanced Academic Performance	0%	28%	43%	29%	0%

Section 6: Opportunities to Express Opinions

Category	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Express your opinions about student performance levels	0%	7%	36%	57%	0%
Ask questions about the standards and how they will be used	0%	0%	50%	50%	0%
Ask questions about the process of making cut score recommendations	0%	0%	43%	57%	0%
Interact with your fellow panelists	0%	0%	21%	79%	0%

Section 7: Respect

Party	No	Sometimes	Yes	Omit
Fellow panelists	0%	0%	79%	21%
Facilitators	0%	0%	100%	0%

SCIENCE — CHEMISTRY

A total of 12 panelists responded to the process evaluation survey.

Section 1: Meeting Success

Meeting Component	Not Successful	Partially Successful	Successful	Very Successful	Omit
Introduction to the process of setting performance standards	0%	0%	17%	83%	0%
Discussion of the performance labels and the definitions	0%	0%	17%	83%	0%
Taking the actual assessment(s)	0%	0%	25%	75%	0%
Overview of the item mapping procedure	0%	0%	17%	83%	0%
Practice exercise for the item-mapping procedure	0%	0%	8%	92%	0%
Feedback data provided in each round	0%	0%	17%	83%	0%
Discussion after each round	0%	0%	8%	92%	0%
Articulation	0%	0%	59%	33%	8%

Section 2: Usefulness of Activities and Information

Activity or Information	Not Useful	Somewhat Useful	Useful	Very Useful	Omit
Specific Performance Level Descriptors (PLDs)	0%	0%	50%	50%	0%
Training in the bookmark standard setting method	0%	8%	17%	75%	0%
Feedback data provided after Round 1	0%	0%	8%	92%	0%
Feedback data provided after Round 2	0%	0%	17%	83%	0%
Presentation of data across courses	0%	0%	33%	67%	0%

Section 3: Adequacy of Meeting Elements

Meeting Element	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Training provided	0%	0%	25%	75%	0%
Amount of time spent training	0%	0%	50%	50%	0%
Feedback provided between rounds	0%	0%	25%	75%	0%

Meeting Element	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Facilities used for the session	0%	0%	8%	92%	0%
Total amount of time in breakout groups to make judgments	0%	0%	50%	50%	0%
Number of rounds for the judgments	0%	0%	58%	42%	0%

Section 4: Specific PLDs

Performance Category	Not Confident	Somewhat Confident	Confident	Very Confident	Omit
Level I: Unsatisfactory Academic Performance	0%	0%	50%	50%	0%
Level II: Satisfactory Academic Performance	0%	8%	42%	50%	0%
Level III: Advanced Academic Performance	0%	8%	42%	50%	0%

Section 5: Cut Score Recommendations

Cut Score	Not Confident	Somewhat Confident	Confident	Very Confident	Omit
Level II: Satisfactory Academic Performance	0%	8%	58%	33%	0%
Level III: Advanced Academic Performance	0%	0%	50%	50%	0%

Section 6: Opportunities to Express Opinions

Category	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Express your opinions about student performance levels	0%	0%	25%	75%	0%
Ask questions about the standards and how they will be used	0%	0%	25%	67%	8%
Ask questions about the process of making cut score recommendations	0%	0%	17%	83%	0%
Interact with your fellow panelists	0%	0%	0%	100%	0%

Section 7: Respect

Party	No	Sometimes	Yes	Omit
Fellow panelists	0%	0%	100%	0%
Facilitators	0%	0%	100%	0%

SCIENCE — PHYSICS

A total of 14 panelists responded to the process evaluation survey.

Section 1: Meeting Success

Meeting Component	Not Successful	Partially Successful	Successful	Very Successful	Omit
Introduction to the process of setting performance standards	0%	7%	57%	36%	0%
Discussion of the performance labels and the definitions	0%	7%	50%	43%	0%
Taking the actual assessment(s)	0%	0%	29%	71%	0%
Overview of the item mapping procedure	0%	0%	43%	57%	0%

Meeting Component	Not Successful	Partially Successful	Successful	Very Successful	Omit
Practice exercise for the item-mapping procedure	0%	0%	43%	57%	0%
Feedback data provided in each round	0%	0%	21%	79%	0%
Discussion after each round	0%	7%	22%	57%	14%
Articulation	0%	14%	50%	36%	0%

Section 2: Usefulness of Activities and Information

Activity or Information	Not Useful	Somewhat Useful	Useful	Very Useful	Omit
Specific Performance Level Descriptors (PLDs)	0%	0%	29%	71%	0%
Training in the bookmark standard setting method	0%	0%	36%	57%	7%
Feedback data provided after Round 1	0%	0%	21%	79%	0%
Feedback data provided after Round 2	0%	0%	21%	79%	0%
Presentation of data across courses	0%	7%	29%	64%	0%

Section 3: Adequacy of Meeting Elements

Meeting Element	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Training provided	0%	0%	43%	50%	7%
Amount of time spent training	0%	0%	36%	64%	0%
Feedback provided between rounds	0%	0%	43%	57%	0%
Facilities used for the session	0%	0%	29%	71%	0%
Total amount of time in breakout groups to make judgments	0%	7%	43%	50%	0%
Number of rounds for the judgments	0%	29%	50%	21%	0%

Section 4: Specific PLDs

Performance Category	Not Confident	Somewhat Confident	Confident	Very Confident	Omit
Level I: Unsatisfactory Academic Performance	0%	21%	36%	43%	0%
Level II: Satisfactory Academic Performance	0%	21%	36%	43%	0%
Level III: Advanced Academic Performance	7%	0%	57%	36%	0%

Section 5: Cut Score Recommendations

Cut Score	Not Confident	Somewhat Confident	Confident	Very Confident	Omit
Level II: Satisfactory Academic Performance	21%	29%	21%	29%	0%
Level III: Advanced Academic Performance	14%	22%	43%	21%	0%

Section 6: Opportunities to Express Opinions

Category	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Express your opinions about student performance levels	0%	0%	50%	50%	0%
Ask questions about the standards and how they will be used	0%	0%	29%	71%	0%
Ask questions about the process of making cut score recommendations	0%	7%	29%	64%	0%
Interact with your fellow panelists	0%	0%	21%	79%	0%

Section 7: Respect

Party	No	Sometimes	Yes	Omit
Fellow panelists	0%	29%	64%	7%
Facilitators	0%	0%	100%	0%

SOCIAL STUDIES — WORLD GEOGRAPHY

A total of 14 panelists responded to the process evaluation survey.

Section 1: Meeting Success

Meeting Component	Not Successful	Partially Successful	Successful	Very Successful	Omit
Introduction to the process of setting performance standards	0%	7%	50%	36%	7%
Discussion of the performance labels and the definitions	0%	0%	43%	50%	7%
Taking the actual assessment(s)	0%	0%	29%	64%	7%
Overview of the item mapping procedure	0%	0%	50%	50%	0%
Practice exercise for the item-mapping procedure	0%	0%	43%	50%	7%
Feedback data provided in each round	0%	0%	36%	64%	0%
Discussion after each round	0%	0%	29%	64%	7%
Articulation	0%	14%	36%	43%	7%

Section 2: Usefulness of Activities and Information

Activity or Information	Not Useful	Somewhat Useful	Useful	Very Useful	Omit
Specific Performance Level Descriptors (PLDs)	0%	0%	36%	64%	0%
Training in the bookmark standard setting method	0%	0%	43%	57%	0%
Feedback data provided after Round 1	0%	0%	21%	79%	0%
Feedback data provided after Round 2	0%	0%	29%	71%	0%
Presentation of data across courses	14%	0%	22%	64%	0%

Section 3: Adequacy of Meeting Elements

Meeting Element	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Training provided	0%	0%	57%	36%	7%
Amount of time spent training	0%	0%	50%	50%	0%
Feedback provided between rounds	0%	0%	43%	50%	7%

Meeting Element	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Facilities used for the session	0%	0%	21%	79%	0%
Total amount of time in breakout groups to make judgments	0%	0%	43%	57%	0%
Number of rounds for the judgments	0%	0%	64%	36%	0%

Section 4: Specific PLDs

Performance Category	Not Confident	Somewhat Confident	Confident	Very Confident	Omit
Level I: Unsatisfactory Academic Performance	0%	0%	22%	71%	7%
Level II: Satisfactory Academic Performance	0%	0%	36%	64%	0%
Level III: Advanced Academic Performance	0%	0%	36%	64%	0%

Section 5: Cut Score Recommendations

Cut Score	Not Confident	Somewhat Confident	Confident	Very Confident	Omit
Level II: Satisfactory Academic Performance	0%	0%	43%	57%	0%
Level III: Advanced Academic Performance	0%	0%	43%	57%	0%

Section 6: Opportunities to Express Opinions

Category	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Express your opinions about student performance levels	0%	7%	36%	57%	0%
Ask questions about the standards and how they will be used	7%	7%	29%	50%	7%
Ask questions about the process of making cut score recommendations	0%	7%	36%	57%	0%
Interact with your fellow panelists	0%	0%	14%	86%	0%

Section 7: Respect

Party	No	Sometimes	Yes	Omit
Fellow panelists	0%	0%	100%	0%
Facilitators	0%	7%	93%	0%

SOCIAL STUDIES — WORLD HISTORY

A total of 13 panelists responded to the process evaluation survey.

Section 1: Meeting Success

Meeting Component	Not Successful	Partially Successful	Successful	Very Successful	Omit
Introduction to the process of setting performance standards	0%	0%	30%	70%	0%
Discussion of the performance labels and the definitions	0%	0%	40%	60%	0%
Taking the actual assessment(s)	0%	0%	10%	90%	0%
Overview of the item mapping procedure	0%	20%	10%	60%	10%

Meeting Component	Not Successful	Partially Successful	Successful	Very Successful	Omit
Practice exercise for the item-mapping procedure	0%	20%	10%	70%	0%
Feedback data provided in each round	10%	10%	20%	60%	0%
Discussion after each round	0%	0%	10%	90%	0%
Articulation	0%	0%	20%	80%	0%

Section 2: Usefulness of Activities and Information

Activity or Information	Not Useful	Somewhat Useful	Useful	Very Useful	Omit
Specific Performance Level Descriptors (PLDs)	0%	0%	40%	60%	0%
Training in the bookmark standard setting method	0%	0%	30%	70%	0%
Feedback data provided after Round 1	0%	0%	40%	60%	0%
Feedback data provided after Round 2	0%	0%	50%	50%	0%
Presentation of data across courses	0%	0%	50%	50%	0%

Section 3: Adequacy of Meeting Elements

Meeting Element	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Training provided	0%	10%	10%	80%	0%
Amount of time spent training	0%	10%	10%	80%	0%
Feedback provided between rounds	0%	10%	30%	60%	0%
Facilities used for the session	0%	0%	10%	90%	0%
Total amount of time in breakout groups to make judgments	0%	0%	20%	80%	0%
Number of rounds for the judgments	0%	0%	20%	80%	0%

Section 4: Specific PLDs

Performance Category	Not Confident	Somewhat Confident	Confident	Very Confident	Omit
Level I: Unsatisfactory Academic Performance	0%	10%	40%	40%	10%
Level II: Satisfactory Academic Performance	0%	10%	40%	50%	0%
Level III: Advanced Academic Performance	0%	10%	30%	60%	0%

Section 5: Cut Score Recommendations

Cut Score	Not Confident	Somewhat Confident	Confident	Very Confident	Omit
Level II: Satisfactory Academic Performance	0%	30%	30%	40%	0%
Level III: Advanced Academic Performance	0%	20%	30%	50%	0%

Section 6: Opportunities to Express Opinions

Category	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Express your opinions about student performance levels	0%	0%	10%	90%	0%
Ask questions about the standards and how they will be used	0%	0%	10%	80%	10%
Ask questions about the process of making cut score recommendations	0%	0%	10%	90%	0%
Interact with your fellow panelists	0%	10%	0%	90%	0%

Section 7: Respect

Party	No	Sometimes	Yes	Omit
Fellow panelists	0%	0%	90%	10%
Facilitators	0%	0%	100%	0%

SOCIAL STUDIES — U.S. HISTORY

A total of 12 panelists responded to the process evaluation survey.

Section 1: Meeting Success

Meeting Component	Not Successful	Partially Successful	Successful	Very Successful	Omit
Introduction to the process of setting performance standards	8%	0%	42%	50%	0%
Discussion of the performance labels and the definitions	0%	0%	34%	58%	8%
Taking the actual assessment(s)	0%	0%	25%	75%	0%
Overview of the item mapping procedure	0%	0%	25%	75%	0%
Practice exercise for the item-mapping procedure	0%	0%	33%	67%	0%
Feedback data provided in each round	0%	0%	42%	58%	0%
Discussion after each round	0%	0%	42%	58%	0%
Articulation	0%	16%	42%	42%	0%

Section 2: Usefulness of Activities and Information

Activity or Information	Not Useful	Somewhat Useful	Useful	Very Useful	Omit
Specific Performance Level Descriptors (PLDs)	0%	0%	17%	83%	0%
Training in the bookmark standard setting method	0%	0%	33%	67%	0%
Feedback data provided after Round 1	0%	0%	17%	83%	0%
Feedback data provided after Round 2	0%	0%	17%	83%	0%
Presentation of data across courses	9%	8%	25%	58%	0%

Section 3: Adequacy of Meeting Elements

Meeting Element	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Training provided	0%	0%	25%	67%	8%
Amount of time spent training	0%	9%	33%	58%	0%
Feedback provided between rounds	0%	0%	33%	67%	0%

Meeting Element	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Facilities used for the session	0%	0%	8%	92%	0%
Total amount of time in breakout groups to make judgments	0%	0%	25%	75%	0%
Number of rounds for the judgments	0%	0%	42%	58%	0%

Section 4: Specific PLDs

Performance Category	Not Confident	Somewhat Confident	Confident	Very Confident	Omit
Level I: Unsatisfactory Academic Performance	0%	17%	25%	58%	0%
Level II: Satisfactory Academic Performance	0%	17%	25%	58%	0%
Level III: Advanced Academic Performance	0%	16%	42%	42%	0%

Section 5: Cut Score Recommendations

Cut Score	Not Confident	Somewhat Confident	Confident	Very Confident	Omit
Level II: Satisfactory Academic Performance	0%	25%	33%	42%	0%
Level III: Advanced Academic Performance	0%	25%	42%	33%	0%

Section 6: Opportunities to Express Opinions

Category	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Express your opinions about student performance levels	0%	0%	25%	75%	0%
Ask questions about the standards and how they will be used	0%	0%	50%	50%	0%
Ask questions about the process of making cut score recommendations	0%	0%	33%	67%	0%
Interact with your fellow panelists	0%	0%	0%	100%	0%

Section 7: Respect

Party	No	Sometimes	Yes	Omit
Fellow panelists	0%	0%	92%	8%
Facilitators	0%	0%	100%	0%

GRADE 8 MATHEMATICS

A total of 13 panelists responded to the process evaluation survey.

Section 1: Meeting Success

Meeting Component	Not Successful	Partially Successful	Successful	Very Successful	Omit
Introduction to the process of setting performance standards	0%	8%	46%	46%	0%
Discussion of the performance labels and the definitions	0%	0%	54%	46%	0%
Taking the actual assessment(s)	0%	8%	46%	46%	0%
Overview of the item mapping procedure	0%	0%	54%	46%	0%
Practice exercise for the item-mapping procedure	0%	0%	46%	54%	0%

Feedback data provided in each round	0%	0%	46%	54%	0%
Discussion after each round	0%	15%	38%	46%	0%

Section 2: Usefulness of Activities and Information

Activity or Information	Not Useful	Somewhat Useful	Useful	Very Useful	Omit
Specific Performance Level Descriptors (PLDs)	0%	8%	23%	69%	0%
Training in the bookmark standard setting method	0%	0%	31%	69%	0%
Feedback data provided after Round 1	0%	0%	46%	54%	0%
Feedback data provided after Round 2	0%	8%	38%	54%	0%

Section 3: Adequacy of Meeting Elements

Meeting Element	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Training provided	0%	0%	31%	69%	0%
Amount of time spent training	0%	0%	38%	62%	0%
Feedback provided between rounds	0%	0%	54%	46%	0%
Facilities used for the session	0%	0%	15%	85%	0%
Total amount of time in breakout groups to make judgments	0%	0%	54%	46%	0%
Number of rounds for the judgments	0%	0%	46%	54%	0%

Section 4: Specific PLDs

Performance Category	Not Confident	Somewhat Confident	Confident	Very Confident	Omit
Level I: Unsatisfactory Academic Performance	0%	8%	38%	54%	0%
Level II: Satisfactory Academic Performance	0%	8%	46%	46%	0%
Level III: Advanced Academic Performance	0%	0%	46%	54%	0%

Section 5: Cut Score Recommendations

Cut Score	Not Confident	Somewhat Confident	Confident	Very Confident	Omit
Level II: Satisfactory Academic Performance	8%	15%	38%	38%	0%
Level III: Advanced Academic Performance	0%	0%	31%	69%	0%

Section 6: Opportunities to Express Opinions

Category	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Express your opinions about student performance levels	0%	8%	38%	54%	0%
Ask questions about the standards and how they will be used	0%	0%	54%	46%	0%
Ask questions about the process of making cut score recommendations	0%	8%	38%	54%	0%
Interact with your fellow panelists	0%	8%	54%	38%	0%

Section 7: Respect

Party	No	Sometimes	Yes	Omit
Fellow panelists	0%	15%	85%	0%
Facilitators	0%	0%	100%	0%

GRADE 8 READING

A total of 14 panelists responded to the process evaluation survey.

Section 1: Meeting Success

Meeting Component	Not Successful	Partially Successful	Successful	Very Successful	Omit
Introduction to the process of setting performance standards	0%	0%	14%	86%	0%
Discussion of the performance labels and the definitions	0%	0%	14%	86%	0%
Taking the actual assessment(s)	0%	0%	7%	93%	0%
Overview of the item mapping procedure	0%	0%	14%	86%	0%
Practice exercise for the item-mapping procedure	0%	0%	29%	71%	0%
Feedback data provided in each round	0%	0%	21%	79%	0%
Discussion after each round	0%	0%	14%	86%	0%

Section 2: Usefulness of Activities and Information

Activity or Information	Not Useful	Somewhat Useful	Useful	Very Useful	Omit
Specific Performance Level Descriptors (PLDs)	0%	7%	0%	86%	7%
Training in the bookmark standard setting method	0%	0%	14%	86%	0%
Feedback data provided after Round 1	0%	0%	14%	86%	0%
Feedback data provided after Round 2	7%	0%	0%	93%	0%

Section 3: Adequacy of Meeting Elements

Meeting Element	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Training provided	0%	0%	50%	50%	0%
Amount of time spent training	0%	0%	57%	43%	0%
Feedback provided between rounds	0%	7%	14%	79%	0%
Facilities used for the session	0%	0%	21%	71%	7%
Total amount of time in breakout groups to make judgments	0%	7%	29%	64%	0%
Number of rounds for the judgments	0%	7%	43%	50%	0%

Section 4: Specific PLDs

Performance Category	Not Confident	Somewhat Confident	Confident	Very Confident	Omit
Level I: Unsatisfactory Academic Performance	0%	0%	43%	57%	0%
Level II: Satisfactory Academic Performance	0%	7%	29%	64%	0%
Level III: Advanced Academic Performance	0%	0%	50%	50%	0%

Section 5: Cut Score Recommendations

Cut Score	Not Confident	Somewhat Confident	Confident	Very Confident	Omit
Level II: Satisfactory Academic Performance	0%	7%	7%	86%	0%
Level III: Advanced Academic Performance	0%	0%	71%	29%	0%

Section 6: Opportunities to Express Opinions

Category	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Express your opinions about student performance levels	0%	7%	14%	79%	0%
Ask questions about the standards and how they will be used	0%	7%	7%	79%	7%
Ask questions about the process of making cut score recommendations	0%	7%	14%	79%	0%
Interact with your fellow panelists	0%	7%	21%	71%	0%

Section 7: Respect

Party	No	Sometimes	Yes	Omit
Fellow panelists	0%	7%	93%	0%
Facilitators	0%	7%	93%	0%

GRADE 8 SCIENCE

A total of 12 panelists responded to the process evaluation survey.

Section 1: Meeting Success

Meeting Component	Not Successful	Partially Successful	Successful	Very Successful	Omit
Introduction to the process of setting performance standards	0%	8%	58%	33%	0%
Discussion of the performance labels and the definitions	0%	0%	58%	42%	0%
Taking the actual assessment(s)	0%	0%	17%	67%	17%
Overview of the item mapping procedure	0%	0%	50%	50%	0%
Practice exercise for the item-mapping procedure	0%	8%	33%	58%	0%
Feedback data provided in each round	0%	0%	25%	75%	0%
Discussion after each round	0%	0%	17%	83%	0%

Section 2: Usefulness of Activities and Information

Activity or Information	Not Useful	Somewhat Useful	Useful	Very Useful	Omit
Specific Performance Level Descriptors (PLDs)	0%	8%	50%	42%	0%
Training in the bookmark standard setting method	0%	0%	50%	50%	0%
Feedback data provided after Round 1	0%	0%	25%	75%	0%
Feedback data provided after Round 2	0%	0%	25%	75%	0%

Section 3: Adequacy of Meeting Elements

Meeting Element	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Training provided	0%	8%	42%	50%	0%
Amount of time spent training	0%	8%	50%	42%	0%
Feedback provided between rounds	0%	0%	50%	50%	0%
Facilities used for the session	0%	0%	8%	75%	17%
Total amount of time in breakout groups to make judgments	8%	8%	58%	25%	0%
Number of rounds for the judgments	0%	0%	50%	50%	0%

Section 4: Specific PLDs

Performance Category	Not Confident	Somewhat Confident	Confident	Very Confident	Omit
Level I: Unsatisfactory Academic Performance	0%	8%	92%	0%	0%
Level II: Satisfactory Academic Performance	0%	8%	75%	17%	0%
Level III: Advanced Academic Performance	0%	17%	58%	25%	0%

Section 5: Cut Score Recommendations

Cut Score	Not Confident	Somewhat Confident	Confident	Very Confident	Omit
Level II: Satisfactory Academic Performance	0%	0%	67%	33%	0%
Level III: Advanced Academic Performance	0%	8%	58%	33%	0%

Section 6: Opportunities to Express Opinions

Category	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Express your opinions about student performance levels	0%	0%	50%	50%	0%
Ask questions about the standards and how they will be used	0%	0%	33%	58%	8%
Ask questions about the process of making cut score recommendations	0%	0%	50%	50%	0%
Interact with your fellow panelists	0%	8%	33%	58%	0%

Section 7: Respect

Party	No	Sometimes	Yes	Omit
Fellow panelists	0%	0%	100%	0%
Facilitators	0%	0%	100%	0%

GRADE 8 SOCIAL STUDIES

A total of 12 panelists responded to the process evaluation survey.

Section 1: Meeting Success

Meeting Component	Not Successful	Partially Successful	Successful	Very Successful	Omit
Introduction to the process of setting performance standards	0%	0%	25%	75%	0%

Meeting Component	Not Successful	Partially Successful	Successful	Very Successful	Omit
Discussion of the performance labels and the definitions	0%	0%	17%	83%	0%
Taking the actual assessment(s)	0%	0%	8%	83%	8%
Overview of the item mapping procedure	0%	0%	8%	92%	0%
Practice exercise for the item-mapping procedure	0%	0%	17%	75%	8%
Feedback data provided in each round	0%	0%	8%	92%	0%
Discussion after each round	0%	0%	17%	83%	0%

Section 2: Usefulness of Activities and Information

Activity or Information	Not Useful	Somewhat Useful	Useful	Very Useful	Omit
Specific Performance Level Descriptors (PLDs)	0%	8%	25%	67%	0%
Training in the bookmark standard setting method	0%	0%	0%	100%	0%
Feedback data provided after Round 1	0%	0%	17%	83%	0%
Feedback data provided after Round 2	0%	0%	25%	75%	0%

Section 3: Adequacy of Meeting Elements

Meeting Element	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Training provided	0%	0%	17%	83%	0%
Amount of time spent training	0%	0%	17%	83%	0%
Feedback provided between rounds	0%	0%	17%	75%	8%
Facilities used for the session	0%	0%	0%	75%	25%
Total amount of time in breakout groups to make judgments	0%	0%	0%	100%	0%
Number of rounds for the judgments	0%	0%	8%	83%	8%

Section 4: Specific PLDs

Performance Category	Not Confident	Somewhat Confident	Confident	Very Confident	Omit
Level I: Unsatisfactory Academic Performance	8%	8%	42%	42%	0%
Level II: Satisfactory Academic Performance	0%	17%	25%	58%	0%
Level III: Advanced Academic Performance	0%	8%	25%	67%	0%

Section 5: Cut Score Recommendations

Cut Score	Not Confident	Somewhat Confident	Confident	Very Confident	Omit
Level II: Satisfactory Academic Performance	8%	33%	33%	25%	0%
Level III: Advanced Academic Performance	0%	8%	25%	67%	0%

Section 6: Opportunities to Express Opinions

Category	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Express your opinions about student performance levels	0%	0%	8%	92%	0%
Ask questions about the standards and how they will be used	0%	0%	0%	92%	8%

Category	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Ask questions about the process of making cut score recommendations	0%	0%	0%	100%	0%
Interact with your fellow panelists	0%	0%	8%	92%	0%

Section 7: Respect

Party	No	Sometimes	Yes	Omit
Fellow panelists	0%	8%	92%	0%
Facilitators	0%	0%	100%	0%

GRADE 7 WRITING (OCTOBER 2012)

A total of 13 panelists responded to the process evaluation survey.

Section 1: Meeting Success

Meeting Component	Not Successful	Partially Successful	Successful	Very Successful	Omit
Introduction to the process of setting performance standards	0%	0%	23%	77%	0%
Discussion of the performance labels and the definitions	0%	0%	23%	77%	0%
Taking the actual assessment(s)	0%	0%	15%	77%	8%
Overview of the item mapping procedure	0%	0%	15%	85%	0%
Practice exercise for the item-mapping procedure	0%	8%	15%	77%	0%
Feedback data provided in each round	0%	0%	8%	92%	0%
Discussion after each round	0%	0%	8%	85%	8%

Section 2: Usefulness of Activities and Information

Activity or Information	Not Useful	Somewhat Useful	Useful	Very Useful	Omit
Specific Performance Level Descriptors (PLDs)	0%	0%	8%	92%	0%
Training in the bookmark standard setting method	0%	0%	8%	92%	0%
Feedback data provided after Round 1	0%	0%	8%	92%	0%
Feedback data provided after Round 2	0%	0%	8%	92%	0%

Section 3: Adequacy of Meeting Elements

Meeting Element	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Training provided	0%	0%	8%	92%	0%
Amount of time spent training	0%	0%	23%	77%	0%
Feedback provided between rounds	0%	0%	15%	77%	8%
Facilities used for the session	0%	0%	0%	92%	8%
Total amount of time in breakout groups to make judgments	0%	0%	15%	85%	0%
Number of rounds for the judgments	0%	0%	15%	85%	0%

Section 4: Specific PLDs

Performance Category	Not Confident	Somewhat Confident	Confident	Very Confident	Omit
Level I: Unsatisfactory Academic Performance	0%	0%	15%	85%	0%
Level II: Satisfactory Academic Performance	0%	0%	15%	77%	8%
Level III: Advanced Academic Performance	0%	0%	23%	77%	0%

Section 5: Cut Score Recommendations

Cut Score	Not Confident	Somewhat Confident	Confident	Very Confident	Omit
Level II: Satisfactory Academic Performance	0%	0%	31%	69%	0%
Level III: Advanced Academic Performance	0%	0%	38%	62%	0%

Section 6: Opportunities to Express Opinions

Category	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Express your opinions about student performance levels	0%	0%	0%	100%	0%
Ask questions about the standards and how they will be used	0%	0%	8%	92%	0%
Ask questions about the process of making cut score recommendations	0%	0%	0%	100%	0%
Interact with your fellow panelists	0%	0%	8%	92%	0%

Section 7: Respect

Party	No	Sometimes	Yes	Omit
Fellow panelists	0%	0%	100%	0%
Facilitators	0%	0%	100%	0%

GRADE 7 WRITING (NOVEMBER 2012)

A total of 10 panelists responded to the process evaluation survey.

Section 1: Meeting Success

Meeting Component	Not Successful	Partially Successful	Successful	Very Successful	Omit
Introduction to the process of setting performance standards	0%	0%	0%	100%	0%
Discussion of the performance labels and the definitions	0%	0%	0%	100%	0%
Taking the actual assessment(s)	0%	0%	0%	100%	0%
Overview of the item mapping procedure	0%	0%	0%	100%	0%

Section 2: Usefulness of Activities and Information

Activity or Information	Not Useful	Somewhat Useful	Useful	Very Useful	Omit
Specific Performance Level Descriptors (PLDs)	0%	0%	20%	80%	0%
Training in the bookmark standard setting method	0%	0%	0%	100%	0%
Feedback data provided after Round 1	0%	0%	0%	100%	0%

Section 3: Adequacy of Meeting Elements

Meeting Element	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Training provided	0%	0%	10%	90%	0%
Amount of time spent training	0%	0%	0%	100%	0%
Feedback provided between rounds	0%	0%	10%	90%	0%

Section 4: Specific PLDs

Performance Category	Not Confident	Somewhat Confident	Confident	Very Confident	Omit
Level I: Unsatisfactory Academic Performance	0%	0%	40%	60%	0%
Level II: Satisfactory Academic Performance	0%	0%	20%	80%	0%

Section 6: Opportunities to Express Opinions

Category	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Express your opinions about student performance levels	0%	0%	0%	100%	0%
Ask questions about the standards and how they will be used	0%	0%	0%	100%	0%
Ask questions about the process of making cut score recommendations	0%	0%	0%	100%	0%

Section 7: Respect

Party	No	Sometimes	Yes	Omit
Fellow panelists	0%	0%	100%	0%
Facilitators	0%	0%	100%	0%

GRADES 6 AND 7 MATHEMATICS

A total of 15 panelists responded to the process evaluation survey.

Section 1: Meeting Success

Meeting Component	Not Successful	Partially Successful	Successful	Very Successful	Omit
Introduction to the process of setting performance standards	0%	0%	7%	93%	0%
Discussion of the performance labels and the definitions	0%	0%	0%	100%	0%
Taking the actual assessment(s)	0%	0%	13%	87%	0%
Overview of the item mapping procedure	0%	0%	7%	93%	0%
Practice exercise for the item-mapping procedure	0%	0%	13%	87%	0%
Feedback data provided in each round	0%	0%	13%	87%	0%
Discussion after each round	0%	0%	13%	87%	0%

Section 2: Usefulness of Activities and Information

Activity or Information	Not Useful	Somewhat Useful	Useful	Very Useful	Omit
Specific Performance Level Descriptors (PLDs)	0%	13%	27%	60%	0%
Training in the bookmark standard setting method	0%	7%	7%	87%	0%
Feedback data provided after Round 1	0%	0%	7%	93%	0%
Feedback data provided after Round 2	0%	0%	7%	93%	0%

Section 3: Adequacy of Meeting Elements

Meeting Element	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Training provided	0%	0%	7%	93%	0%
Amount of time spent training	0%	0%	27%	67%	7%
Feedback provided between rounds	0%	0%	7%	93%	0%
Facilities used for the session	0%	0%	0%	100%	0%
Total amount of time in breakout groups to make judgments	0%	0%	20%	80%	0%
Number of rounds for the judgments	0%	0%	27%	73%	0%

Section 4: Specific PLDs

Performance Category	Not Confident	Somewhat Confident	Confident	Very Confident	Omit
Level I: Unsatisfactory Academic Performance	0%	7%	47%	40%	7%
Level II: Satisfactory Academic Performance	0%	0%	60%	40%	0%
Level III: Advanced Academic Performance	0%	7%	60%	27%	7%

Section 5: Cut Score Recommendations

Cut Score	Not Confident	Somewhat Confident	Confident	Very Confident	Omit
Level II: Satisfactory Academic Performance	0%	7%	60%	33%	0%
Level III: Advanced Academic Performance	0%	7%	47%	40%	7%

Section 6: Opportunities to Express Opinions

Category	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Express your opinions about student performance levels	0%	7%	13%	80%	0%
Ask questions about the standards and how they will be used	0%	0%	20%	73%	7%
Ask questions about the process of making cut score recommendations	0%	0%	7%	87%	7%
Interact with your fellow panelists	7%	0%	0%	93%	0%

Section 7: Respect

Party	No	Sometimes	Yes	Omit
Fellow panelists	7%	0%	93%	0%
Facilitators	0%	0%	100%	0%

GRADES 6 AND 7 READING

A total of 16 panelists responded to the process evaluation survey.

Section 1: Meeting Success

Meeting Component	Not Successful	Partially Successful	Successful	Very Successful	Omit
Introduction to the process of setting performance standards	0%	0%	50%	50%	0%
Discussion of the performance labels and the definitions	0%	6%	17%	78%	0%
Taking the actual assessment(s)	0%	0%	6%	94%	0%
Overview of the item mapping procedure	0%	0%	22%	78%	0%
Practice exercise for the item-mapping procedure	0%	0%	28%	72%	0%
Feedback data provided in each round	0%	0%	28%	72%	0%
Discussion after each round	0%	0%	33%	67%	0%

Section 2: Usefulness of Activities and Information

Activity or Information	Not Useful	Somewhat Useful	Useful	Very Useful	Omit
Specific Performance Level Descriptors (PLDs)	0%	0%	28%	72%	0%
Training in the bookmark standard setting method	0%	0%	17%	83%	0%
Feedback data provided after Round 1	0%	0%	11%	89%	0%
Feedback data provided after Round 2	0%	0%	11%	89%	0%

Section 3: Adequacy of Meeting Elements

Meeting Element	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Training provided	0%	0%	22%	78%	0%
Amount of time spent training	0%	0%	33%	67%	0%
Feedback provided between rounds	0%	0%	11%	89%	0%
Facilities used for the session	0%	0%	0%	94%	6%
Total amount of time in breakout groups to make judgments	0%	6%	17%	78%	0%
Number of rounds for the judgments	0%	0%	17%	83%	0%

Section 4: Specific PLDs

Performance Category	Not Confident	Somewhat Confident	Confident	Very Confident	Omit
Level I: Unsatisfactory Academic Performance	0%	0%	22%	78%	0%
Level II: Satisfactory Academic Performance	0%	0%	28%	72%	0%
Level III: Advanced Academic Performance	0%	0%	28%	72%	0%

Section 5: Cut Score Recommendations

Cut Score	Not Confident	Somewhat Confident	Confident	Very Confident	Omit
Level II: Satisfactory Academic Performance	0%	0%	56%	44%	0%
Level III: Advanced Academic Performance	0%	0%	56%	39%	6%

Section 6: Opportunities to Express Opinions

Category	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Express your opinions about student performance levels	0%	0%	22%	78%	0%
Ask questions about the standards and how they will be used	0%	0%	17%	83%	0%
Ask questions about the process of making cut score recommendations	0%	0%	11%	89%	0%
Interact with your fellow panelists	0%	0%	11%	89%	0%

Section 7: Respect

Party	No	Sometimes	Yes	Omit
Fellow panelists	0%	6%	94%	0%
Facilitators	0%	0%	94%	6%

GRADE 5 MATHEMATICS

A total of 12 panelists responded to the process evaluation survey.

Section 1: Meeting Success

Meeting Component	Not Successful	Partially Successful	Successful	Very Successful	Omit
Introduction to the process of setting performance standards	0%	17%	75%	8%	0%
Discussion of the performance labels and the definitions	0%	8%	75%	17%	0%
Taking the actual assessment(s)	0%	0%	50%	50%	0%
Overview of the item mapping procedure	0%	8%	67%	25%	0%
Practice exercise for the item-mapping procedure	0%	25%	33%	42%	0%
Feedback data provided in each round	0%	8%	50%	42%	0%
Discussion after each round	0%	25%	33%	42%	0%

Section 2: Usefulness of Activities and Information

Activity or Information	Not Useful	Somewhat Useful	Useful	Very Useful	Omit
Specific Performance Level Descriptors (PLDs)	0%	33%	33%	33%	0%
Training in the bookmark standard setting method	0%	17%	50%	33%	0%
Feedback data provided after Round 1	0%	8%	33%	58%	0%
Feedback data provided after Round 2	0%	8%	25%	67%	0%

Section 3: Adequacy of Meeting Elements

Meeting Element	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Training provided	0%	8%	58%	33%	0%
Amount of time spent training	0%	17%	50%	33%	0%
Feedback provided between rounds	0%	8%	50%	42%	0%
Facilities used for the session	0%	0%	25%	75%	0%
Total amount of time in breakout groups to make judgments	0%	25%	25%	50%	0%
Number of rounds for the judgments	0%	0%	67%	33%	0%

Section 4: Specific PLDs

Performance Category	Not Confident	Somewhat Confident	Confident	Very Confident	Omit
Level I: Unsatisfactory Academic Performance	8%	0%	67%	17%	8%
Level II: Satisfactory Academic Performance	8%	0%	67%	25%	0%
Level III: Advanced Academic Performance	0%	17%	58%	25%	0%

Section 5: Cut Score Recommendations

Cut Score	Not Confident	Somewhat Confident	Confident	Very Confident	Omit
Level II: Satisfactory Academic Performance	8%	17%	58%	17%	0%
Level III: Advanced Academic Performance	0%	17%	67%	17%	0%

Section 6: Opportunities to Express Opinions

Category	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Express your opinions about student performance levels	0%	0%	50%	50%	0%
Ask questions about the standards and how they will be used	0%	8%	42%	50%	0%
Ask questions about the process of making cut score recommendations	0%	8%	42%	50%	0%
Interact with your fellow panelists	0%	0%	58%	42%	0%

Section 7: Respect

Party	No	Sometimes	Yes	Omit
Fellow panelists	0%	8%	92%	0%
Facilitators	0%	0%	100%	0%

GRADE 5 ENGLISH READING

A total of 14 panelists responded to the process evaluation survey.

Section 1: Meeting Success

Meeting Component	Not Successful	Partially Successful	Successful	Very Successful	Omit
Introduction to the process of setting performance standards	0%	0%	14%	68%	0%
Discussion of the performance labels and the definitions	0%	0%	21%	79%	0%
Taking the actual assessment(s)	0%	0%	0%	100%	0%
Overview of the item mapping procedure	0%	0%	21%	79%	0%
Practice exercise for the item-mapping procedure	0%	0%	14%	86%	0%
Feedback data provided in each round	0%	0%	0%	100%	0%
Discussion after each round	0%	7%	7%	86%	0%

Section 2: Usefulness of Activities and Information

Activity or Information	Not Useful	Somewhat Useful	Useful	Very Useful	Omit
Specific Performance Level Descriptors (PLDs)	0%	0%	7%	93%	0%
Training in the bookmark standard setting method	0%	0%	0%	100%	0%
Feedback data provided after Round 1	0%	0%	7%	93%	0%
Feedback data provided after Round 2	0%	0%	7%	93%	0%

Section 3: Adequacy of Meeting Elements

Meeting Element	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Training provided	0%	0%	7%	93%	0%
Amount of time spent training	0%	0%	21%	79%	0%
Feedback provided between rounds	0%	0%	21%	79%	0%
Facilities used for the session	0%	0%	0%	86%	14%
Total amount of time in breakout groups to make judgments	0%	0%	43%	57%	0%
Number of rounds for the judgments	0%	0%	50%	43%	7%

Section 4: Specific PLDs

Performance Category	Not Confident	Somewhat Confident	Confident	Very Confident	Omit
Level I: Unsatisfactory Academic Performance	0%	0%	21%	79%	0%
Level II: Satisfactory Academic Performance	0%	0%	21%	79%	0%
Level III: Advanced Academic Performance	0%	0%	36%	64%	0%

Section 5: Cut Score Recommendations

Cut Score	Not Confident	Somewhat Confident	Confident	Very Confident	Omit
Level II: Satisfactory Academic Performance	0%	7%	43%	50%	0%
Level III: Advanced Academic Performance	0%	0%	43%	57%	0%

Section 6: Opportunities to Express Opinions

Category	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Express your opinions about student performance levels	0%	0%	0%	100%	0%
Ask questions about the standards and how they will be used	0%	0%	0%	100%	0%
Ask questions about the process of making cut score recommendations	0%	0%	0%	100%	0%
Interact with your fellow panelists	0%	7%	0%	93%	0%

Section 7: Respect

Party	No	Sometimes	Yes	Omit
Fellow panelists	0%	0%	100%	0%
Facilitators	0%	0%	100%	0%

GRADE 5 SPANISH READING

A total of 12 panelists responded to the process evaluation survey.

Section 1: Meeting Success

Meeting Component	Not Successful	Partially Successful	Successful	Very Successful	Omit
Introduction to the process of setting performance standards	0%	9%	45%	45%	0%
Discussion of the performance labels and the definitions	0%	27%	36%	36%	0%
Taking the actual assessment(s)	0%	0%	45%	55%	0%
Overview of the item mapping procedure	0%	9%	55%	36%	0%
Practice exercise for the item-mapping procedure	0%	18%	45%	36%	0%
Feedback data provided in each round	0%	0%	45%	55%	0%
Discussion after each round	0%	9%	27%	64%	0%

Section 2: Usefulness of Activities and Information

Activity or Information	Not Useful	Somewhat Useful	Useful	Very Useful	Omit
Specific Performance Level Descriptors (PLDs)	0%	18%	27%	55%	0%
Training in the bookmark standard setting method	0%	18%	18%	64%	0%
Feedback data provided after Round 1	0%	18%	27%	55%	0%
Feedback data provided after Round 2	0%	9%	36%	55%	0%

Section 3: Adequacy of Meeting Elements

Meeting Element	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Training provided	0%	9%	64%	27%	0%
Amount of time spent training	9%	9%	55%	27%	0%
Feedback provided between rounds	0%	0%	73%	27%	0%
Facilities used for the session	0%	9%	36%	45%	9%
Total amount of time in breakout groups to make judgments	0%	0%	45%	55%	0%
Number of rounds for the judgments	0%	0%	64%	36%	0%

Section 4: Specific PLDs

Performance Category	Not Confident	Somewhat Confident	Confident	Very Confident	Omit
Level I: Unsatisfactory Academic Performance	0%	18%	64%	18%	0%
Level II: Satisfactory Academic Performance	0%	18%	82%	0%	0%
Level III: Advanced Academic Performance	0%	18%	55%	27%	0%

Section 5: Cut Score Recommendations

Cut Score	Not Confident	Somewhat Confident	Confident	Very Confident	Omit
Level II: Satisfactory Academic Performance	0%	27%	55%	18%	0%
Level III: Advanced Academic Performance	0%	36%	36%	27%	0%

Section 6: Opportunities to Express Opinions

Category	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Express your opinions about student performance levels	0%	0%	45%	55%	0%
Ask questions about the standards and how they will be used	0%	0%	45%	55%	0%
Ask questions about the process of making cut score recommendations	0%	0%	36%	64%	0%
Interact with your fellow panelists	0%	0%	18%	82%	0%

Section 7: Respect

Party	No	Sometimes	Yes	Omit
Fellow panelists	0%	0%	100%	0%
Facilitators	0%	0%	100%	0%

GRADE 5 SCIENCE

A total of 11 panelists responded to the process evaluation survey.

Section 1: Meeting Success

Meeting Component	Not Successful	Partially Successful	Successful	Very Successful	Omit
Introduction to the process of setting performance standards	0%	8%	67%	25%	0%
Discussion of the performance labels and the definitions	0%	8%	67%	25%	0%
Taking the actual assessment(s)	0%	0%	33%	67%	0%
Overview of the item mapping procedure	0%	8%	58%	33%	0%
Practice exercise for the item-mapping procedure	0%	0%	67%	33%	0%
Feedback data provided in each round	0%	0%	17%	83%	0%
Discussion after each round	0%	0%	25%	75%	0%

Section 2: Usefulness of Activities and Information

Activity or Information	Not Useful	Somewhat Useful	Useful	Very Useful	Omit
Specific Performance Level Descriptors (PLDs)	0%	0%	42%	58%	0%
Training in the bookmark standard setting method	0%	8%	25%	58%	8%
Feedback data provided after Round 1	0%	0%	25%	75%	0%
Feedback data provided after Round 2	0%	0%	25%	75%	0%

Section 3: Adequacy of Meeting Elements

Meeting Element	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Training provided	0%	8%	58%	33%	0%
Amount of time spent training	0%	8%	50%	42%	0%
Feedback provided between rounds	0%	0%	42%	58%	0%
Facilities used for the session	0%	0%	17%	67%	17%
Total amount of time in breakout groups to make judgments	0%	0%	50%	50%	0%
Number of rounds for the judgments	0%	8%	50%	42%	0%

Section 4: Specific PLDs

Performance Category	Not Confident	Somewhat Confident	Confident	Very Confident	Omit
Level I: Unsatisfactory Academic Performance	0%	25%	50%	25%	0%
Level II: Satisfactory Academic Performance	0%	25%	42%	25%	8%
Level III: Advanced Academic Performance	0%	33%	17%	42%	8%

Section 5: Cut Score Recommendations

Cut Score	Not Confident	Somewhat Confident	Confident	Very Confident	Omit
Level II: Satisfactory Academic Performance	8%	33%	25%	33%	0%
Level III: Advanced Academic Performance	8%	17%	25%	50%	0%

Section 6: Opportunities to Express Opinions

Category	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Express your opinions about student performance levels	0%	17%	33%	50%	0%
Ask questions about the standards and how they will be used	0%	0%	33%	50%	17%
Ask questions about the process of making cut score recommendations	0%	17%	42%	42%	0%
Interact with your fellow panelists	0%	0%	58%	42%	0%

Section 7: Respect

Party	No	Sometimes	Yes	Omit
Fellow panelists	0%	8%	92%	0%
Facilitators	0%	8%	92%	0%

GRADE 4 ENGLISH WRITING (OCTOBER 2012)

A total of 11 panelists responded to the process evaluation survey.

Section 1: Meeting Success

Meeting Component	Not Successful	Partially Successful	Successful	Very Successful	Omit
Introduction to the process of setting performance standards	0%	0%	27%	73%	0%
Discussion of the performance labels and the definitions	0%	9%	0%	91%	0%
Taking the actual assessment(s)	0%	0%	0%	100%	0%
Overview of the item mapping procedure	0%	0%	27%	73%	0%
Practice exercise for the item-mapping procedure	0%	0%	36%	64%	0%
Feedback data provided in each round	0%	0%	36%	64%	0%
Discussion after each round	0%	0%	18%	73%	9%

Section 2: Usefulness of Activities and Information

Activity or Information	Not Useful	Somewhat Useful	Useful	Very Useful	Omit
Specific Performance Level Descriptors (PLDs)	0%	0%	27%	73%	0%
Training in the bookmark standard setting method	0%	0%	18%	82%	0%
Feedback data provided after Round 1	0%	0%	18%	82%	0%
Feedback data provided after Round 2	0%	0%	18%	82%	0%

Section 3: Adequacy of Meeting Elements

Meeting Element	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Training provided	0%	0%	36%	64%	0%
Amount of time spent training	0%	0%	36%	55%	9%
Feedback provided between rounds	0%	9%	27%	64%	0%
Facilities used for the session	0%	0%	0%	91%	9%
Total amount of time in breakout groups to make judgments	0%	0%	18%	82%	0%
Number of rounds for the judgments	0%	0%	18%	82%	0%

Section 4: Specific PLDs

Performance Category	Not Confident	Somewhat Confident	Confident	Very Confident	Omit
Level I: Unsatisfactory Academic Performance	0%	9%	36%	45%	9%
Level II: Satisfactory Academic Performance	0%	0%	45%	55%	0%
Level III: Advanced Academic Performance	0%	0%	45%	55%	0%

Section 5: Cut Score Recommendations

Cut Score	Not Confident	Somewhat Confident	Confident	Very Confident	Omit
Level II: Satisfactory Academic Performance	0%	9%	36%	55%	0%
Level III: Advanced Academic Performance	0%	0%	55%	45%	0%

Section 6: Opportunities to Express Opinions

Category	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Express your opinions about student performance levels	0%	0%	27%	73%	0%
Ask questions about the standards and how they will be used	0%	0%	18%	64%	18%
Ask questions about the process of making cut score recommendations	0%	0%	27%	73%	0%
Interact with your fellow panelists	0%	0%	18%	82%	0%

Section 7: Respect

Party	No	Sometimes	Yes	Omit
Fellow panelists	0%	0%	100%	0%
Facilitators	0%	0%	100%	0%

GRADE 4 ENGLISH WRITING (NOVEMBER 2012)

A total of 10 panelists responded to the process evaluation survey.

Section 1: Meeting Success

Meeting Component	Not Successful	Partially Successful	Successful	Very Successful	Omit
Introduction to the process of setting performance standards	0%	0%	0%	100%	0%
Discussion of the performance labels and the definitions	0%	0%	0%	100%	0%
Taking the actual assessment(s)	0%	0%	0%	100%	0%
Overview of the item mapping procedure	0%	0%	10%	90%	0%

Section 2: Usefulness of Activities and Information

Activity or Information	Not Useful	Somewhat Useful	Useful	Very Useful	Omit
Specific Performance Level Descriptors (PLDs)	0%	0%	10%	90%	0%
Training in the bookmark standard setting method	0%	0%	0%	100%	0%
Feedback data provided after Round 1	0%	0%	0%	100%	0%

Section 3: Adequacy of Meeting Elements

Meeting Element	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Training provided	0%	0%	0%	100%	0%
Amount of time spent training	0%	0%	0%	100%	0%
Feedback provided between rounds	0%	0%	0%	100%	0%

Section 4: Specific PLDs

Performance Category	Not Confident	Somewhat Confident	Confident	Very Confident	Omit
Level I: Unsatisfactory Academic Performance	0%	0%	10%	90%	0%
Level II: Satisfactory Academic Performance	0%	0%	20%	80%	0%

Section 6: Opportunities to Express Opinions

Category	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Express your opinions about student performance levels	0%	0%	0%	90%	10%
Ask questions about the standards and how they will be used	0%	0%	0%	90%	10%
Ask questions about the process of making cut score recommendations	0%	0%	10%	80%	10%

Section 7: Respect

Party	No	Sometimes	Yes	Omit
Fellow panelists	0%	0%	90%	10%
Facilitators	0%	0%	90%	10%

GRADE 4 SPANISH WRITING (OCTOBER 2012)

A total of 13 panelists responded to the process evaluation survey.

Section 1: Meeting Success

Meeting Component	Not Successful	Partially Successful	Successful	Very Successful	Omit
Introduction to the process of setting performance standards	0%	8%	8%	85%	0%
Discussion of the performance labels and the definitions	0%	0%	8%	92%	0%
Taking the actual assessment(s)	0%	0%	8%	92%	0%
Overview of the item mapping procedure	0%	8%	8%	85%	0%
Practice exercise for the item-mapping procedure	8%	0%	23%	69%	0%
Feedback data provided in each round	0%	0%	8%	92%	0%
Discussion after each round	0%	0%	15%	85%	0%

Section 2: Usefulness of Activities and Information

Activity or Information	Not Useful	Somewhat Useful	Useful	Very Useful	Omit
Specific Performance Level Descriptors (PLDs)	0%	0%	0%	100%	0%
Training in the bookmark standard setting method	0%	15%	15%	69%	0%
Feedback data provided after Round 1	0%	0%	15%	85%	0%
Feedback data provided after Round 2	0%	0%	0%	100%	0%

Section 3: Adequacy of Meeting Elements

Meeting Element	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Training provided	8%	0%	31%	54%	8%
Amount of time spent training	8%	0%	38%	54%	0%
Feedback provided between rounds	0%	0%	46%	46%	8%
Facilities used for the session	0%	0%	23%	69%	8%
Total amount of time in breakout groups to make judgments	0%	8%	46%	46%	0%
Number of rounds for the judgments	0%	0%	77%	23%	0%

Section 4: Specific PLDs

Performance Category	Not Confident	Somewhat Confident	Confident	Very Confident	Omit
Level I: Unsatisfactory Academic Performance	0%	0%	38%	62%	0%
Level II: Satisfactory Academic Performance	0%	0%	38%	62%	0%
Level III: Advanced Academic Performance	0%	0%	38%	62%	0%

Section 5: Cut Score Recommendations

Cut Score	Not Confident	Somewhat Confident	Confident	Very Confident	Omit
Level II: Satisfactory Academic Performance	0%	0%	69%	31%	0%
Level III: Advanced Academic Performance	0%	0%	69%	31%	0%

Section 6: Opportunities to Express Opinions

Category	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Express your opinions about student performance levels	0%	0%	31%	69%	0%
Ask questions about the standards and how they will be used	0%	0%	31%	69%	0%
Ask questions about the process of making cut score recommendations	0%	0%	23%	77%	0%
Interact with your fellow panelists	0%	0%	46%	54%	0%

Section 7: Respect

Party	No	Sometimes	Yes	Omit
Fellow panelists	0%	0%	100%	0%
Facilitators	0%	0%	100%	0%

GRADE 4 SPANISH WRITING (NOVEMBER 2012)

A total of 10 panelists responded to the process evaluation survey.

Section 1: Meeting Success

Meeting Component	Not Successful	Partially Successful	Successful	Very Successful	Omit
Introduction to the process of setting performance standards	0%	0%	10%	90%	0%
Discussion of the performance labels and the definitions	0%	0%	10%	90%	0%
Taking the actual assessment(s)	0%	0%	10%	70%	20%
Overview of the item mapping procedure	0%	0%	10%	90%	0%

Section 2: Usefulness of Activities and Information

Activity or Information	Not Useful	Somewhat Useful	Useful	Very Useful	Omit
Specific Performance Level Descriptors (PLDs)	0%	0%	0%	100%	0%
Training in the bookmark standard setting method	0%	20%	10%	70%	0%
Feedback data provided after Round 1	0%	10%	20%	70%	0%

Section 3: Adequacy of Meeting Elements

Meeting Element	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Training provided	0%	0%	50%	50%	0%
Amount of time spent training	0%	0%	30%	70%	0%
Feedback provided between rounds	0%	0%	50%	50%	0%

Section 4: Specific PLDs

Performance Category	Not Confident	Somewhat Confident	Confident	Very Confident	Omit
Level I: Unsatisfactory Academic Performance	0%	0%	30%	70%	0%
Level II: Satisfactory Academic Performance	0%	0%	40%	60%	0%

Section 6: Opportunities to Express Opinions

Category	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Express your opinions about student performance levels	0%	0%	50%	50%	0%
Ask questions about the standards and how they will be used	0%	0%	30%	70%	0%
Ask questions about the process of making cut score recommendations	0%	0%	20%	80%	0%

Section 7: Respect

Party	No	Sometimes	Yes	Omit
Fellow panelists	0%	0%	100%	0%
Facilitators	0%	0%	100%	0%

GRADES 3 AND 4 MATHEMATICS

A total of 15 panelists responded to the process evaluation survey.

Section 1: Meeting Success

Meeting Component	Not Successful	Partially Successful	Successful	Very Successful	Omit
Introduction to the process of setting performance standards	0%	0%	13%	87%	0%
Discussion of the performance labels and the definitions	0%	0%	20%	80%	0%
Taking the actual assessment(s)	0%	0%	0%	100%	0%
Overview of the item mapping procedure	0%	0%	27%	73%	0%
Practice exercise for the item-mapping procedure	0%	0%	33%	67%	0%
Feedback data provided in each round	0%	0%	20%	80%	0%
Discussion after each round	0%	0%	20%	80%	0%

Section 2: Usefulness of Activities and Information

Activity or Information	Not Useful	Somewhat Useful	Useful	Very Useful	Omit
Specific Performance Level Descriptors (PLDs)	0%	0%	20%	80%	0%
Training in the bookmark standard setting method	0%	0%	33%	67%	0%
Feedback data provided after Round 1	0%	7%	0%	93%	0%
Feedback data provided after Round 2	0%	7%	7%	87%	0%

Section 3: Adequacy of Meeting Elements

Meeting Element	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Training provided	0%	0%	20%	80%	0%
Amount of time spent training	0%	0%	20%	80%	0%
Feedback provided between rounds	0%	0%	27%	73%	0%
Facilities used for the session	0%	0%	0%	100%	0%
Total amount of time in breakout groups to make judgments	0%	0%	13%	87%	0%
Number of rounds for the judgments	0%	0%	27%	73%	0%

Section 4: Specific PLDs

Performance Category	Not Confident	Somewhat Confident	Confident	Very Confident	Omit
Level I: Unsatisfactory Academic Performance	0%	0%	60%	40%	0%
Level II: Satisfactory Academic Performance	0%	7%	60%	33%	0%
Level III: Advanced Academic Performance	0%	0%	53%	47%	0%

Section 5: Cut Score Recommendations

Cut Score	Not Confident	Somewhat Confident	Confident	Very Confident	Omit
Level II: Satisfactory Academic Performance	0%	0%	40%	60%	0%
Level III: Advanced Academic Performance	0%	0%	47%	53%	0%

Section 6: Opportunities to Express Opinions

Category	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Express your opinions about student performance levels	0%	0%	13%	87%	0%
Ask questions about the standards and how they will be used	0%	0%	13%	87%	0%
Ask questions about the process of making cut score recommendations	0%	0%	13%	87%	0%
Interact with your fellow panelists	0%	0%	7%	93%	0%

Section 7: Respect

Party	No	Sometimes	Yes	Omit
Fellow panelists	0%	0%	100%	0%
Facilitators	7%	0%	93%	0%

GRADES 3 AND 4 ENGLISH READING

A total of 16 panelists responded to the process evaluation survey.

Section 1: Meeting Success

Meeting Component	Not Successful	Partially Successful	Successful	Very Successful	Omit
Introduction to the process of setting performance standards	0%	6%	56%	38%	0%
Discussion of the performance labels and the definitions	0%	19%	44%	38%	0%
Taking the actual assessment(s)	0%	0%	25%	75%	0%
Overview of the item mapping procedure	0%	13%	56%	31%	0%
Practice exercise for the item-mapping procedure	0%	6%	44%	50%	0%
Feedback data provided in each round	0%	0%	25%	75%	0%
Discussion after each round	0%	0%	19%	81%	0%

Section 2: Usefulness of Activities and Information

Activity or Information	Not Useful	Somewhat Useful	Useful	Very Useful	Omit
Specific Performance Level Descriptors (PLDs)	0%	6%	38%	56%	0%
Training in the bookmark standard setting method	0%	0%	44%	56%	0%
Feedback data provided after Round 1	0%	0%	25%	75%	0%
Feedback data provided after Round 2	0%	0%	19%	81%	0%

Section 3: Adequacy of Meeting Elements

Meeting Element	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Training provided	0%	0%	56%	44%	0%
Amount of time spent training	0%	6%	50%	44%	0%
Feedback provided between rounds	0%	0%	44%	56%	0%
Facilities used for the session	0%	0%	6%	81%	13%
Total amount of time in breakout groups to make judgments	0%	0%	25%	75%	0%
Number of rounds for the judgments	0%	0%	69%	31%	0%

Section 4: Specific PLDs

Performance Category	Not Confident	Somewhat Confident	Confident	Very Confident	Omit
Level I: Unsatisfactory Academic Performance	0%	13%	50%	38%	0%
Level II: Satisfactory Academic Performance	0%	6%	63%	31%	0%
Level III: Advanced Academic Performance	0%	6%	56%	38%	0%

Section 5: Cut Score Recommendations

Cut Score	Not Confident	Somewhat Confident	Confident	Very Confident	Omit
Level II: Satisfactory Academic Performance	0%	19%	50%	31%	0%
Level III: Advanced Academic Performance	6%	6%	44%	44%	0%

Section 6: Opportunities to Express Opinions

Category	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Express your opinions about student performance levels	0%	0%	38%	63%	0%
Ask questions about the standards and how they will be used	0%	0%	38%	56%	6%
Ask questions about the process of making cut score recommendations	0%	0%	44%	56%	0%
Interact with your fellow panelists	0%	0%	31%	69%	0%

Section 7: Respect

Party	No	Sometimes	Yes	Omit
Fellow panelists	0%	6%	94%	0%
Facilitators	0%	0%	100%	0%

GRADES 3 AND 4 SPANISH READING

A total of 15 panelists responded to the process evaluation survey.

Section 1: Meeting Success

Meeting Component	Not Successful	Partially Successful	Successful	Very Successful	Omit
Introduction to the process of setting performance standards	0%	14%	29%	57%	0%
Discussion of the performance labels and the definitions	0%	21%	21%	57%	0%
Taking the actual assessment(s)	0%	21%	21%	57%	0%
Overview of the item mapping procedure	0%	21%	21%	57%	0%
Practice exercise for the item-mapping procedure	0%	14%	36%	50%	0%
Feedback data provided in each round	0%	21%	29%	50%	0%
Discussion after each round	0%	29%	14%	57%	0%

Section 2: Usefulness of Activities and Information

Activity or Information	Not Useful	Somewhat Useful	Useful	Very Useful	Omit
Specific Performance Level Descriptors (PLDs)	7%	7%	14%	64%	7%
Training in the bookmark standard setting method	0%	14%	21%	57%	7%
Feedback data provided after Round 1	0%	7%	14%	57%	21%
Feedback data provided after Round 2	0%	7%	36%	50%	7%

Section 3: Adequacy of Meeting Elements

Meeting Element	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Training provided	0%	14%	50%	36%	0%
Amount of time spent training	0%	21%	43%	36%	0%
Feedback provided between rounds	0%	21%	36%	43%	0%
Facilities used for the session	0%	7%	29%	50%	14%
Total amount of time in breakout groups to make judgments	0%	7%	50%	43%	0%
Number of rounds for the judgments	0%	7%	43%	50%	0%

Section 4: Specific PLDs

Performance Category	Not Confident	Somewhat Confident	Confident	Very Confident	Omit
Level I: Unsatisfactory Academic Performance	29%	0%	36%	36%	0%
Level II: Satisfactory Academic Performance	29%	0%	36%	36%	0%
Level III: Advanced Academic Performance	21%	7%	29%	43%	0%

Section 5: Cut Score Recommendations

Cut Score	Not Confident	Somewhat Confident	Confident	Very Confident	Omit
Level II: Satisfactory Academic Performance	14%	21%	29%	36%	0%
Level III: Advanced Academic Performance	7%	21%	29%	43%	0%

Section 6: Opportunities to Express Opinions

Category	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate	Omit
Express your opinions about student performance levels	0%	21%	36%	43%	0%
Ask questions about the standards and how they will be used	7%	14%	36%	43%	0%
Ask questions about the process of making cut score recommendations	7%	14%	36%	43%	0%
Interact with your fellow panelists	0%	7%	36%	57%	0%

Section 7: Respect

Party	No	Sometimes	Yes	Omit
Fellow panelists	0%	7%	93%	0%
Facilitators	0%	14%	86%	0%

Appendix 16: Summary of Cut Score Recommendations

This appendix provides a summary of the cut score recommendations (based on the OIB page number) after each judgment round of the standard -setting committee meetings, as well as after the cross-course articulation or group discussion and reasonableness review.

STAAR EOC ASSESSMENTS

ENGLISH

	English I Reading		English II Reading		English III Reading		English I Writing		English II Writing		English III Writing	
	Level II	**Level III**	**Level II**	**Level III**	**Level II**	**Level III**	**Level II**	**Level III**	**Level II**	**Level III**	**Level II**	**Level III**
Round 1	42	54	40	54	42	61	48	60	45	54	49	55
Round 2	41	54	41	54	41	61	48	60	45	54	42	54
Round 3	41	54	41	54	38	57	48	60	45	54	44	54
Articulation	41	54	41	54	41	57	48	60	45	54	46	54
Reasonableness Review	**41**	**54**	**41**	**54**	**41**	**57**	**48**	**60**	**45**	**54**	**46**	**54**

MATHEMATICS

	Algebra I		Algebra II		Geometry	
	Level II	**Level III**	**Level II**	**Level III**	**Level II**	**Level III**
Round 1	20	42	24	45	25	51
Round 2	24	45	23	44	25	49
Round 3	26	47	22	44	25	49
Articulation	26	53	22	44	20	49
Reasonableness Review	**32**	**53**	**22**	**44**	**25**	**49**

Science

	Biology		Chemistry		Physics	
	Level II	**Level III**	**Level II**	**Level III**	**Level II**	**Level III**
Round 1	30	62	29	55	31	50
Round 2	33	60	27	55	26	45
Round 3	31	61	28	55	23	38
Articulation	31	61	28	55	17	42
Reasonableness Review	**31**	**61**	**28**	**55**	**23**	**49**

Social Studies

	World Geography		World History		U.S. History	
	Level II	**Level III**	**Level II**	**Level III**	**Level II**	**Level III**
Round 1	43	70	34	62	29	69
Round 2	41	70	35	65	39	73
Round 3	41	71	33	63	43	73
Articulation	41	71	33	63	43	73
Reasonableness Review	**48**	**71**	**33**	**63**	**43**	**73**

STAAR 3–8 ASSESSMENTS

Mathematics

	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8	
	Level II	**Level III**	**Level II**	**Level III**	**Level II**	**Level III**	**Level II**	**Level III**	**Level II**	**Level III**	**Level II**	**Level III**
Round 1	40	61	48	59	41	60	43	61	53	68	35	69
Round 2	47	61	48	58	43	63	49	61	49	69	35	68
Round 3	47	59	48	58	48	63	44	61	45	69	35	69
Group Discussion	47	59	48	58	48	63	44	61	45	69	35	69
Reasonableness Review	**47**	**56**	**48**	**58**	**48**	**60**	**44**	**61**	**45**	**69**	**35**	**69**

Reading - English

	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8	
	Level II	**Level III**	**Level II**	**Level III**	**Level II**	**Level III**	**Level II**	**Level III**	**Level II**	**Level III**	**Level II**	**Level III**
Round 1	36	45	43	51	44	54	49	59	50	60	43	53
Round 2	37	45	43	51	44	54	50	59	50	60	39	52
Round 3	38	46	40	50	43	56	49	60	47	60	40	52
Group Discussion	38	46	40	50	43	56	49	60	47	60	40	54
Reasonableness Review	**36**	**46**	**40**	**50**	**43**	**56**	**49**	**60**	**47**	**60**	**40**	**54**

Reading - Spanish

	Grade 3		Grade 4		Grade 5	
	Level II	**Level III**	**Level II**	**Level III**	**Level II**	**Level III**
Round 1	41	54	46	60	39	53
Round 2	40	52	45	59	38	53
Round 3	40	52	44	59	37	52
Group Discussion	40	52	44	59	37	52
Reasonableness Review	**40**	**52**	**44**	**59**	**37**	**52**

Science

	Grade 5		Grade 8	
	Level II	**Level III**	**Level II**	**Level III**
Round 1	50	57	46	62
Round 2	47	56	44	63
Round 3	48	57	43	61
Group Discussion	48	57	43	61
Reasonableness Review	**48**	**57**	**43**	**61**

Social Studies

	Grade 8	
	Level II	**Level III**
Round 1	51	66
Round 2	53	66
Round 3	53	66
Group Discussion	53	66
Reasonableness Review	**53**	**66**

Writing

	Grade 4 English		Grade 4 Spanish		Grade 7	
	Level II	**Level III**	**Level II**	**Level III**	**Level II**	**Level III**
Round 1	48	57	40	49	44	53
Round 2	47	57	40	48	43	53
Round 3	44	57	40	48	43	53
Group Discussion	44	57	40	48	43	53
Reasonableness Review	**46**	**57**	**40**	**48**	**43**	**53**

Appendix 17: Summary of Standard-Setting Panelists' Judgments

This appendix provides descriptive statistics — minimum, maximum, mean, standard deviation, and median — of the standard-setting panelists' cut score recommendations (based on the OIB page number) during each judgment round of the committee meetings.

Statistics are given separately for each assessment.

STAAR EOC Assessments

ENGLISH READING

		Round	Minimum	Maximum	Mean	Standard Deviation	Median
English I Reading	Level II	1	39	46	41.8	2.1	42
		2	37	44	41.3	2.0	41
		3	40	42	41.0	0.4	41
	Level III	1	50	57	54.2	1.8	54
		2	41	56	52.6	4.1	54
		3	50	54	53.2	1.6	54
English II Reading	Level II	1	24	42	38.5	4.9	40
		2	38	42	40.4	1.3	41
		3	39	41	40.2	1.0	41
	Level III	1	48	62	53.6	4.1	54
		2	50	58	53.8	2.5	54
		3	50	62	55.4	3.6	54
English III Reading	Level II	1	40	48	42.8	2.3	42
		2	39	43	41.3	1.0	41
		3	36	39	37.7	1.2	38
	Level III	1	53	65	60.6	3.7	61
		2	56	61	59.3	2.5	61
		3	46	64	55.4	4.4	57

ENGLISH WRITING

		Round	Minimum	Maximum	Mean	Standard Deviation	Median
English I Writing	Level II	1	39	54	47.8	3.4	48
		2	47	51	48.3	1.6	48
		3	47	51	48.8	1.5	48
	Level III	1	50	61	58.8	2.6	60
		2	55	60	58.9	1.5	60
		3	57	60	59.0	1.4	60
English II Writing	Level II	1	41	52	45.1	2.4	45
		2	43	47	45.1	0.8	45
		3	43	46	44.9	0.6	45
	Level III	1	51	55	53.2	1.3	54
		2	50	55	53.4	1.3	54
		3	50	54	53.6	1.2	54
English III Writing	Level II	1	39	54	48.4	5.0	49
		2	40	49	43.4	2.9	42
		3	40	49	44.5	2.8	44
	Level III	1	53	56	54.4	0.8	55
		2	52	55	53.9	1.1	54
		3	53	55	53.9	0.9	54

MATHEMATICS

		Round	Minimum	Maximum	Mean	Standard Deviation	Median
Algebra I	Level II	1	15	40	21.7	6.8	20
		2	19	40	25.2	5.7	24
		3	19	40	26.1	5.6	26
	Level III	1	35	64	46.1	7.5	42
		2	40	64	47.8	7.0	45
		3	40	64	48.9	6.9	47
Algebra II	Level II	1	13	43	25.0	8.9	24
		2	15	39	24.2	6.9	23
		3	15	36	23.6	5.9	22
	Level III	1	31	65	46.1	8.4	45
		2	39	65	46.0	6.7	44
		3	37	60	45.7	6.7	44
Geometry	Level II	1	14	54	25.2	8.4	25
		2	15	49	26.2	7.0	25
		3	15	48	25.0	6.9	25
	Level III	1	38	64	50.2	7.4	51
		2	40	64	50.7	6.3	49
		3	40	63	49.9	6.4	49

SCIENCE

		Round	Minimum	Maximum	Mean	Standard Deviation	Median
Biology	Level II	1	23	55	34.3	9.7	30
		2	29	42	33.3	4.3	33
		3	29	42	33.0	4.6	31
	Level III	1	49	67	59.1	6.4	62
		2	51	66	58.8	5.3	60
		3	53	66	60.4	4.3	61
Chemistry	Level II	1	17	53	30.8	9.8	29
		2	22	40	28.0	5.9	27
		3	22	40	29.1	6.2	28
	Level III	1	42	62	55.3	4.8	55
		2	49	60	54.5	2.6	55
		3	53	60	55.2	1.5	55
Physics	Level II	1	7	55	29.5	11.2	31
		2	17	41	25.0	6.9	26
		3	17	41	23.1	6.8	23
	Level III	1	35	63	48.9	9.2	50
		2	21	55	42.8	8.7	45
		3	35	55	40.9	6.5	38

SOCIAL STUDIES

		Round	Minimum	Maximum	Mean	Standard Deviation	Median
World Geography	Level II	1	19	67	41.1	12.1	43
		2	34	42	39.7	2.5	41
		3	36	47	41.1	2.4	41
	Level III	1	57	77	68.4	6.3	70
		2	64	73	69.8	2.6	70
		3	68	73	70.6	1.9	71
World History	Level II	1	19	56	37.8	13.2	34
		2	29	65	38.9	9.4	35
		3	29	48	34.3	5.6	33
	Level III	1	39	73	59.2	11.9	62
		2	52	77	64.0	6.7	65
		3	49	72	63.0	6.3	63
U.S. History	Level II	1	18	62	35.7	15.1	29
		2	26	57	39.6	10.2	39
		3	35	57	43.9	8.4	43
	Level III	1	62	86	68.3	7.1	69
		2	66	81	71.5	4.5	73
		3	68	83	73.6	4.5	73

STAAR 3–8 Assessments

MATHEMATICS

		Round	Minimum	Maximum	Mean	Standard Deviation	Median
Grade 3 Mathematics	Level II	1	34	52	42.9	5.422	40
		2	40	48	45.4	2.9	47
		3	42	51	47.1	1.8	47
	Level III	1	53	62	59.2	3.21	61
		2	53	62	59.6	3.1	61
		3	53	62	58.7	3.1	59
Grade 4 Mathematics	Level II	1	42	58	48.0	4.12	48
		2	46	49	48.1	0.8	48
		3	47	49	48.1	0.5	48
	Level III	1	53	59	58.1	1.6	59
		2	54	59	57.5	1.5	58
		3	53	59	57.7	1.4	58
Grade 5 Mathematics	Level II	1	38	54	43.3	5.3	41
		2	39	53	44.2	4.0	43
		3	41	54	47.4	3.9	48
	Level III	1	56	68	61.2	4.8	60
		2	56	65	61.1	3.3	63
		3	58	65	62.0	2.0	63
Grade 6 Mathematics	Level II	1	36	52	43.6	4.6	43
		2	44	53	47.6	2.7	49
		3	43	50	45.2	2.2	44
	Level III	1	56	65	61.3	2.4	61
		2	60	65	61.2	1.1	61
		3	59	611	60.9	0.5	61
Grade 7 Mathematics	Level II	1	40	55	50.0	5.1	53
		2	43	54	48.7	4.3	49
		3	43	53	46.7	3.5	45
	Level III	1	59	73	68.3	3.5	68
		2	60	71	67.9	3.2	69
		3	60	71	68.1	2.6	69
Grade 8 Mathematics	Level II	1	28	55	36.6	8.5	35
		2	30	52	39.5	7.5	35
		3	32	53	39.4	7.9	35
	Level III	1	63	72	67.8	2.6	69
		2	63	72	68.4	2.2	68
		3	63	72	68.8	2.2	69

READING

		Round	Minimum	Maximum	Mean	Standard Deviation	Median
Grade 3 English Reading	Level II	1	32	41	36.4	1.9	36
		2	35	39	37.0	1.2	37
		3	36	38	37.4	1.0	38
	Level III	1	43	46	44.6	0.8	45
		2	41	46	44.4	1.3	45
		3	45	46	45.8	0.4	46
Grade 4 English Reading	Level II	1	38	48	43.3	2.4	43
		2	41	47	43.3	1.3	43
		3	39	43	40.3	1.2	40
	Level III	1	48	54	51.3	1.5	51
		2	49	53	50.9	1.0	51
		3	49	43	50.3	0.8	50
Grade 5 English Reading	Level II	1	36	48	43.4	3.5	44
		2	40	48	42.9	2.1	44
		3	39	48	42.6	2.3	43
	Level III	1	50	57	54.2	1.7	54
		2	52	56	54.2	1.4	54
		3	52	57	55.6	1.2	56
Grade 6 Reading	Level II	1	41	52	48.0	4.0	49
		2	42	52	48.7	3.1	50
		3	45	51	48.3	2.2	49
	Level III	1	54	61	59.0	1.7	59
		2	57	62	59.1	1.2	59
		3	52	61	59.2	2.1	60
Grade 7 Reading	Level II	1	18	58	48.3	9.1	50
		2	40	55	48.9	3.8	50
		3	44	51	47.8	1.8	47
	Level III	1	57	64	59.6	2.2	60
		2	58	61	59.4	0.9	60
		3	54	64	59.4	2.0	60
Grade 8 Reading	Level II	1	33	50	42.7	4.8	43
		2	36	47	39.9	3.1	39
		3	38	44	40.2	1.8	40
	Level III	1	48	57	52.5	2.5	53
		2	48	56	51.6	2.4	52
		3	49	55	51.5	1.9	52

SPANISH READING

		Round	Minimum	Maximum	Mean	Standard Deviation	Median
Grade 3 Spanish Reading	Level II	1	37	46	41.3	2.7	41
		2	39	46	40.8	2.4	40
		3	39	46	41.4	2.4	40
	Level III	1	49	59	54.2	2.5	54
		2	50	55	52.6	1.4	52
		3	50	54	52.5	1.2	52
Grade 4 Spanish Reading	Level II	1	43	56	47.1	3.6	46
		2	44	50	45.3	1.5	45
		3	41	46	43.4	2.0	44
	Level III	1	57	64	60.0	1.6	60
		2	58	60	59.3	0.6	59
		3	57	60	58.9	0.9	59
Grade 5 Spanish Reading	Level II	1	30	48	40.9	5.8	39
		2	34	47	40.2	5.3	38
		3	34	40	36.8	1.7	37
	Level III	1	50	55	53.0	1.5	53
		2	48	54	52.5	1.8	53
		3	51	54	51.9	1.0	52

SCIENCE

		Round	Minimum	Maximum	Mean	Standard Deviation	Median
Grade 5 Science	Level II	1	42	53	48.7	3.8	50
		2	42	52	46.9	2.8	47
		3	42	50	46.7	3.4	48
	Level III	1	55	60	57.3	1.5	57
		2	51	57	56.0	1.7	56
		3	56	57	56.5	0.5	57
Grade 8 Science	Level II	1	37	56	36.2	5.0	46
		2	39	49	43.5	2.9	44
		3	40	44	42.5	1.6	43
	Level III	1	55	65	60.7	3.5	62
		2	56	63	60.8	2.8	63
		3	57	63	60.9	1.8	61

SOCIAL STUDIES

		Round	Minimum	Maximum	Mean	Standard Deviation	Median
Grade 8 Social Studies	Level II	1	35	56	49.0	6.8	51
		2	42	57	51.3	5.1	53
		3	42	66	53.0	5.6	53
	Level III	1	49	73	65.2	6.4	66
		2	60	72	66.0	3.8	66
		3	60	76	66.5	4.1	66

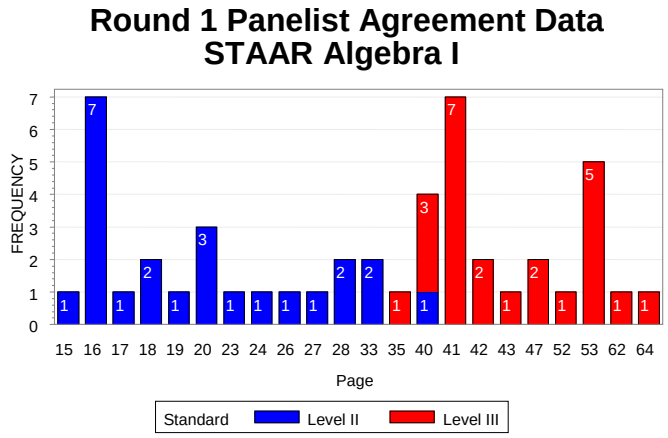
WRITING

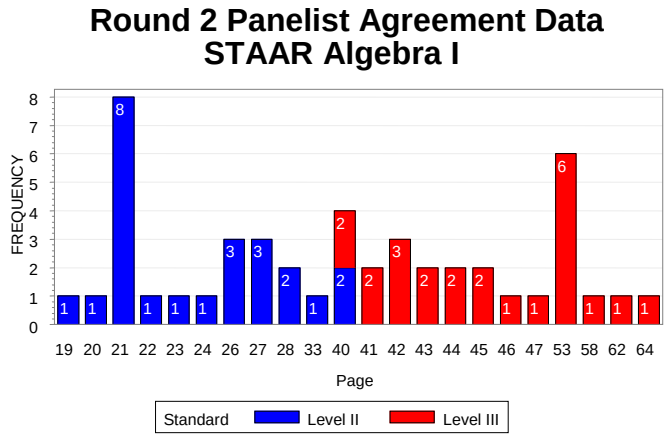
		Round	Minimum	Maximum	Mean	Standard Deviation	Median
Grade 4 English Writing	Level II	1	44	51	47.6	2.8	48
		2	44	48	46.2	1.9	47
		3	44	48	45.2	1.9	44
	Level III	1	56	62	57.8	1.8	57
		2	56	62	57.4	1.7	57
		3	56	62	57.4	1.7	57
Grade 7 Writing	Level II	1	42	48	44.2	1.8	44
		2	42	47	43.2	1.5	43
		3	42	47	43.4	1.4	43
	Level III	1	51	55	52.7	1.5	53
		2	51	54	52.6	1.2	53
		3	51	54	52.8	1.0	53
Grade 4 Spanish Writing	Level II	1	34	45	39.3	3.2	40
		2	37	42	39.2	1.8	40
		3	37	40	39.3	1.2	40
	Level III	1	47	53	49.4	2.4	49
		2	44	53	48.2	2.2	48
		3	47	49	48.1	0.7	48

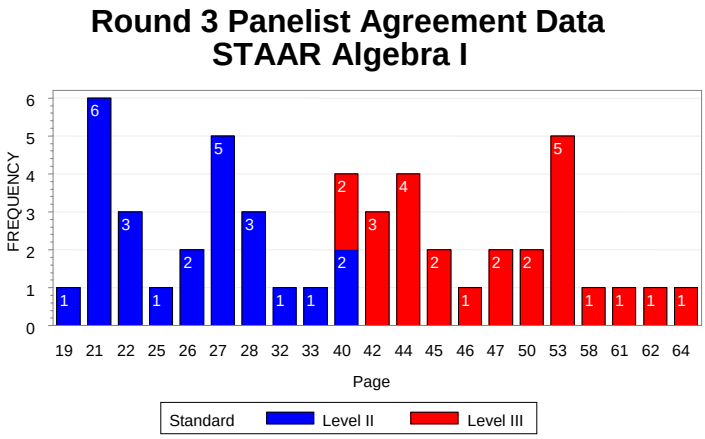
Appendix 18: Standard-Setting Panelists' Agreement Data

This appendix provides the frequency distribution of the recommended cuts (bookmarked page numbers) after each round of judgments for each assessment.

ALGEBRA I

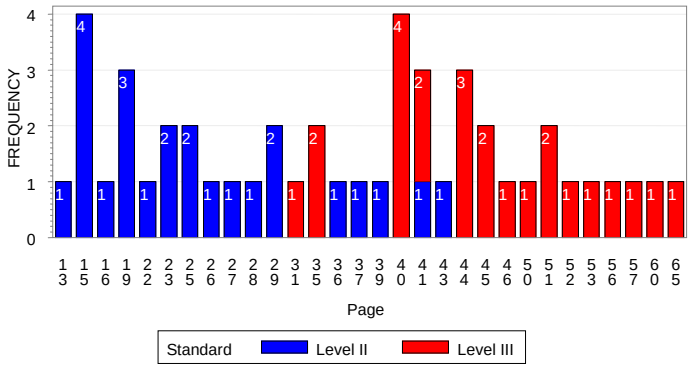






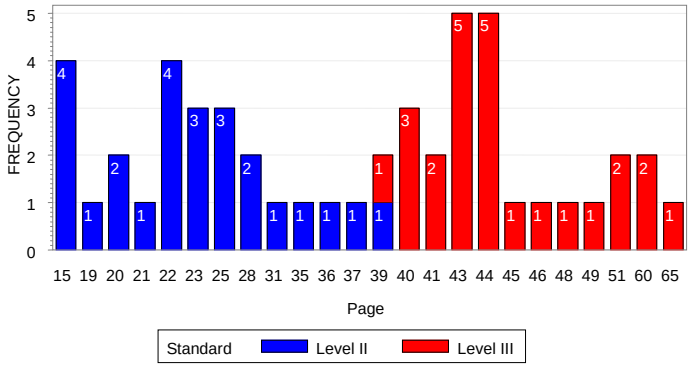
ALGEBRA II

Round 1 Panelist Agreement Data
STAAR Algebra II

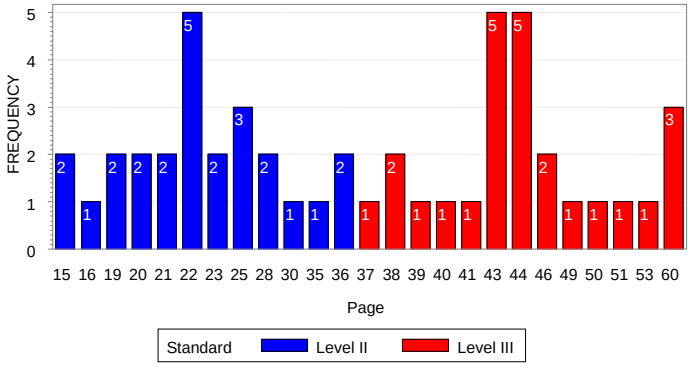


* Numbers along the horizontal axis are stacked due to the large range of panelists' judgments.

Round 2 Panelist Agreement Data
STAAR Algebra II

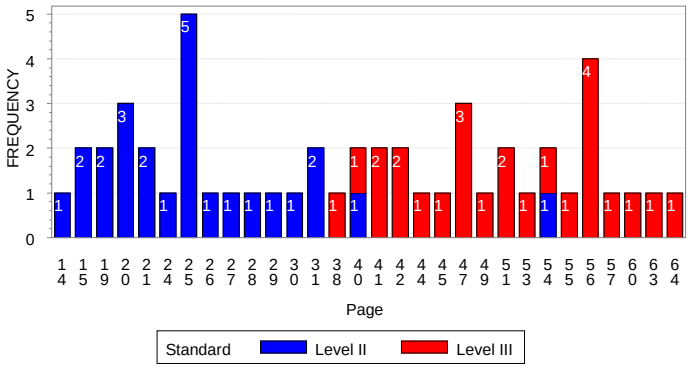


Round 3 Panelist Agreement Data
STAAR Algebra II



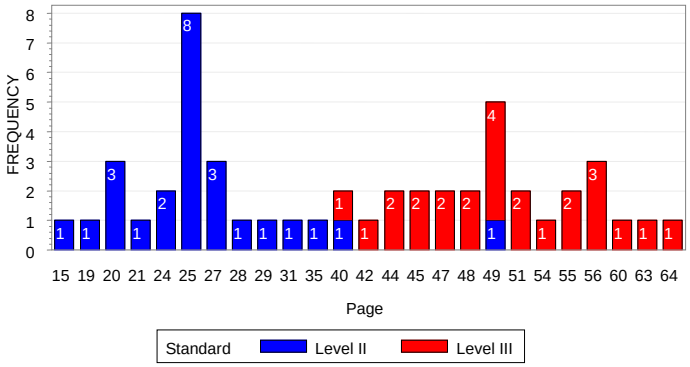
GEOMETRY

Round 1 Panelist Agreement Data
STAAR Geometry

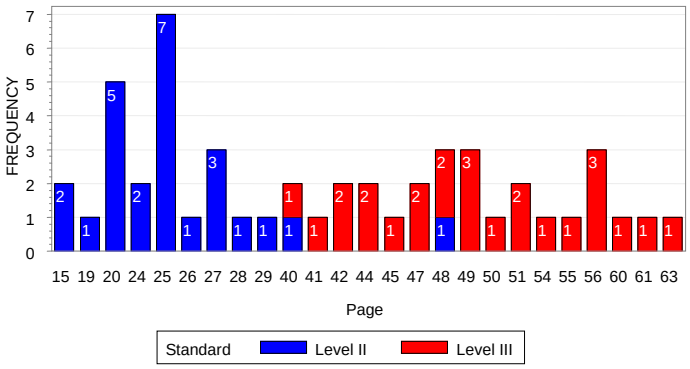


* Numbers along the horizontal axis are stacked due to the large range of panelists' judgments.

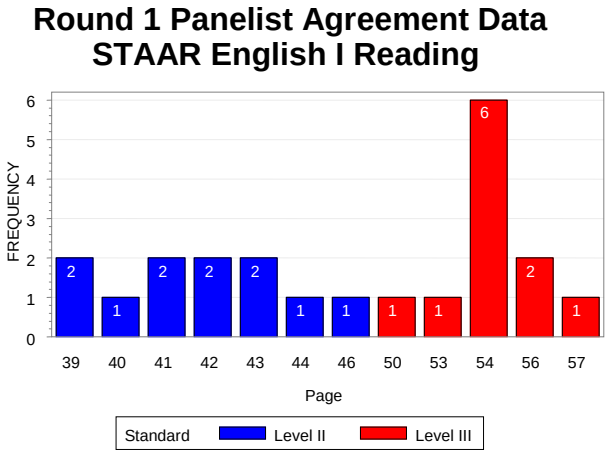
Round 2 Panelist Agreement Data
STAAR Geometry

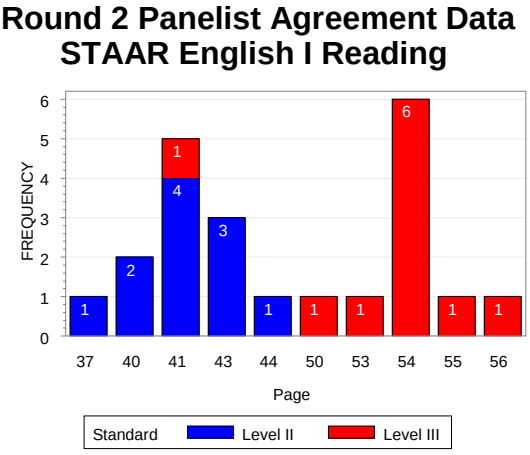


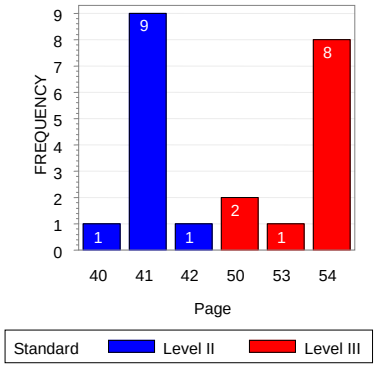
Round 3 Panelist Agreement Data
STAAR Geometry



ENGLISH I READING

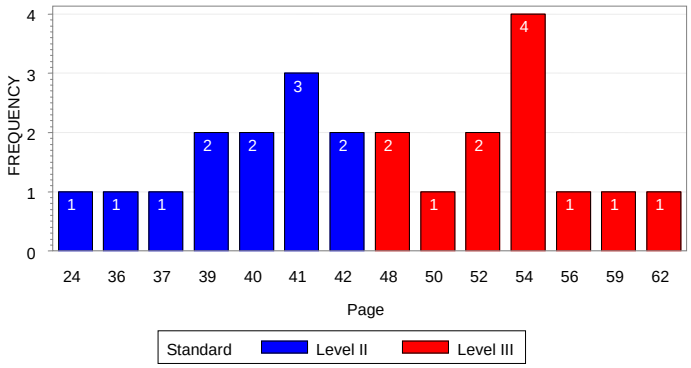




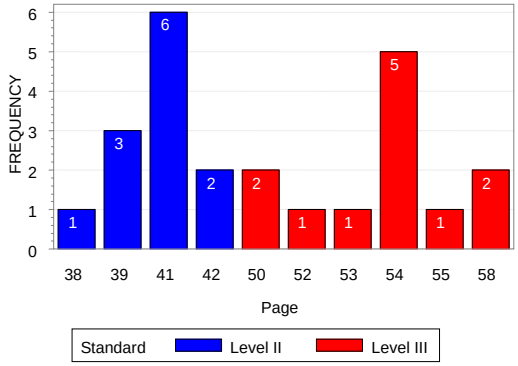


ENGLISH II READING

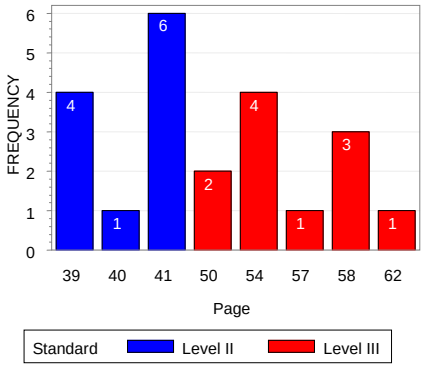
Round 1 Panelist Agreement Data
STAAR English II Reading



Round 2 Panelist Agreement Data
STAAR English II Reading

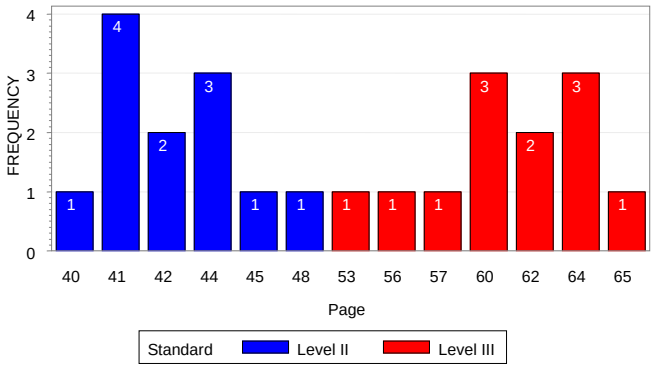


Round 3 Panelist Agreement Data
STAAR English II Reading

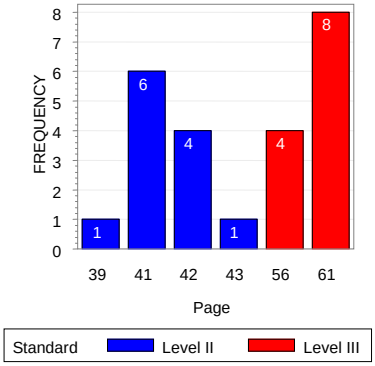


ENGLISH III READING

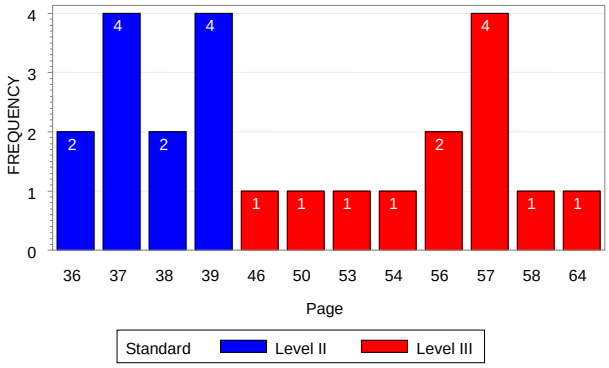
Round 1 Panelist Agreement Data
STAAR English III Reading



Round 2 Panelist Agreement Data
STAAR English III Reading

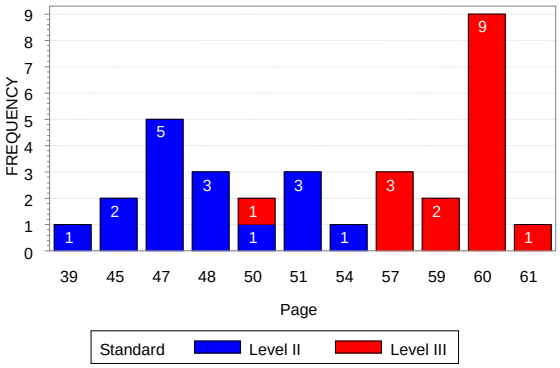


Round 3 Panelist Agreement Data
STAAR English III Reading

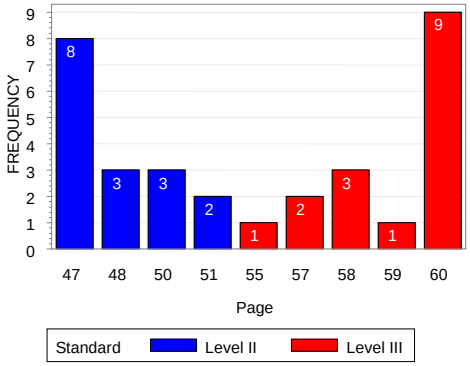


ENGLISH I WRITING

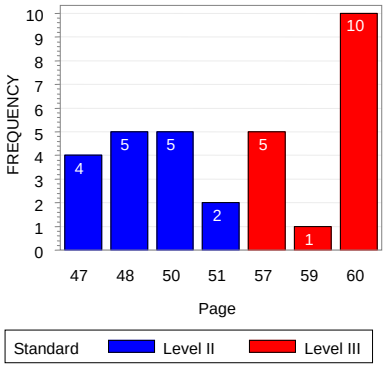
Round 1 Panelist Agreement Data
STAAR English I Writing



Round 2 Panelist Agreement Data
STAAR English I Writing

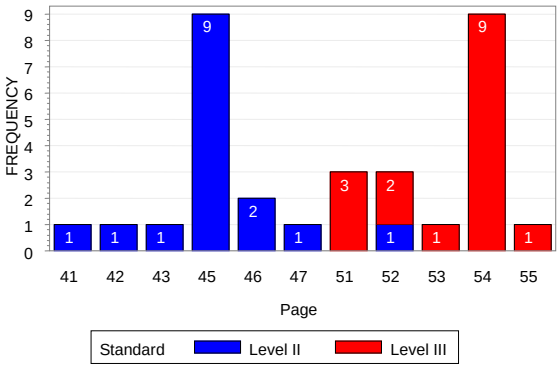


Round 3 Panelist Agreement Data
STAAR English I Writing

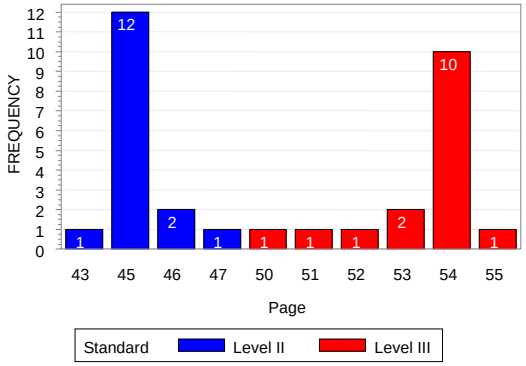


ENGLISH II WRITING

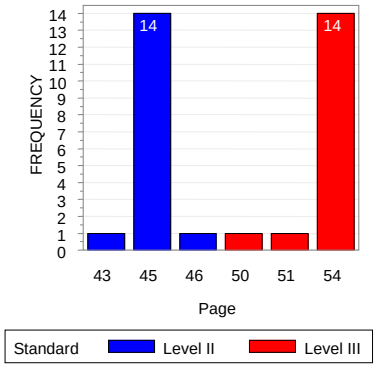
Round 1 Panelist Agreement Data
STAAR English II Writing



Round 2 Panelist Agreement Data
STAAR English II Writing

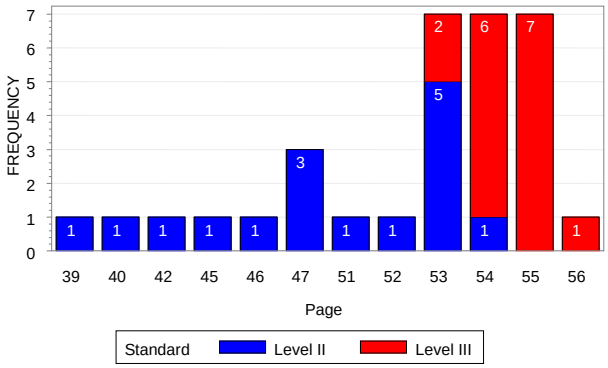


Round 3 Panelist Agreement Data
STAAR English II Writing

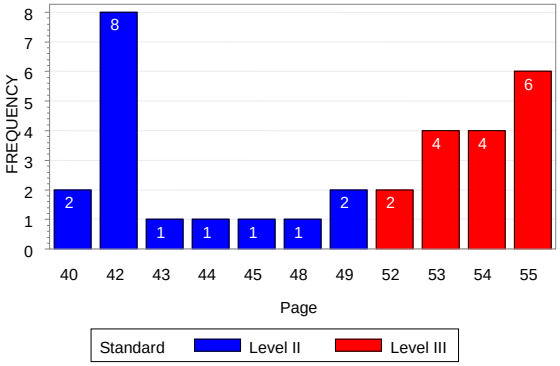


ENGLISH III WRITING

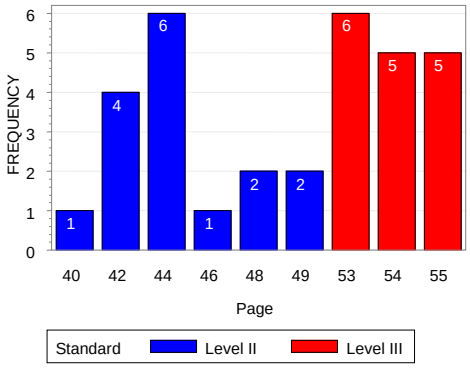
Round 1 Panelist Agreement Data
STAAR English III Writing



Round 2 Panelist Agreement Data
STAAR English III Writing

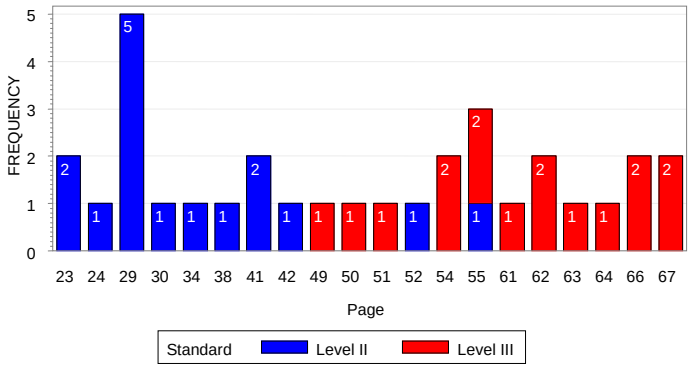


Round 3 Panelist Agreement Data
STAAR English III Writing

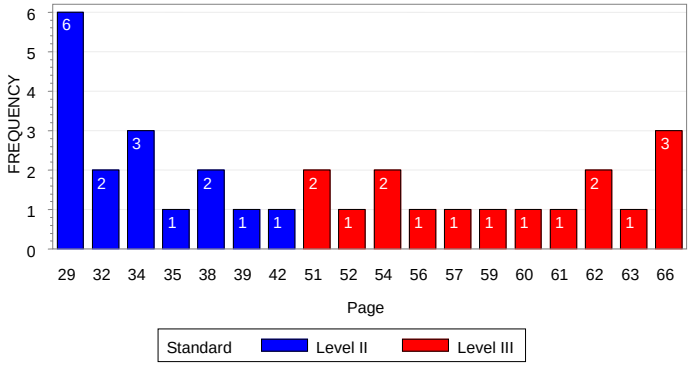


BIOLOGY

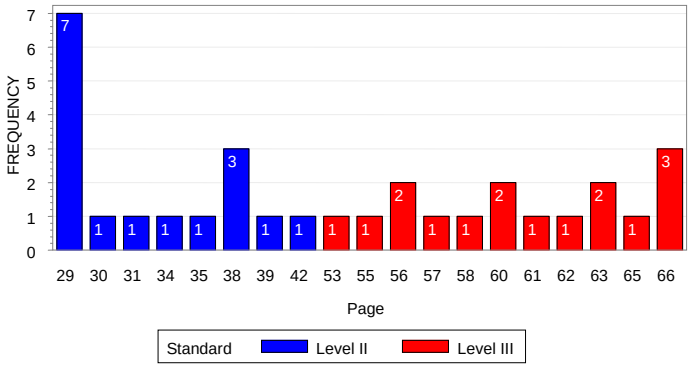
Round 1 Panelist Agreement Data
STAAR Biology



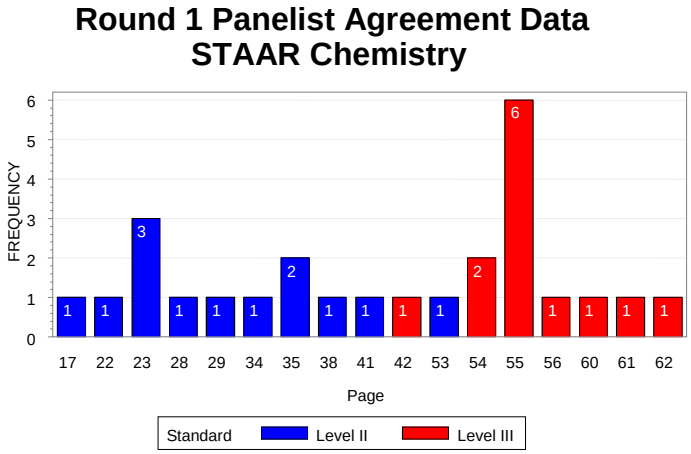
Round 2 Panelist Agreement Data
STAAR Biology

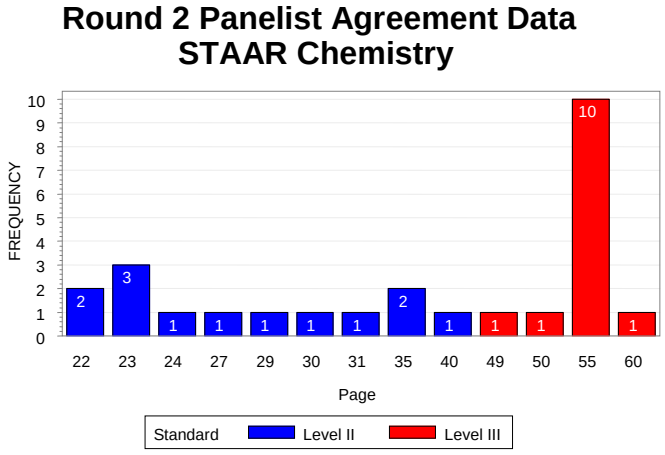


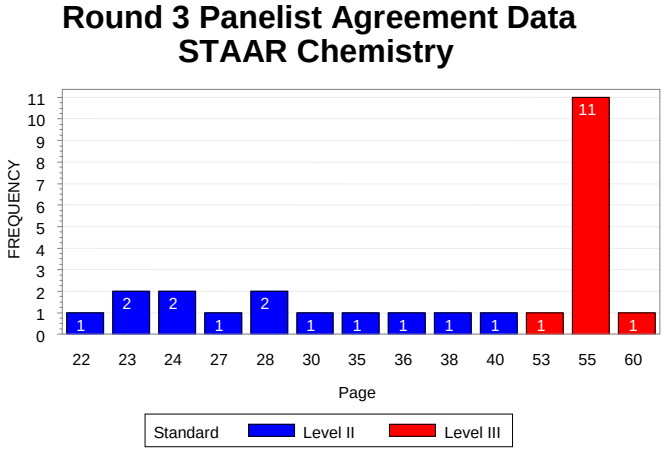
Round 3 Panelist Agreement Data
STAAR Biology



CHEMISTRY

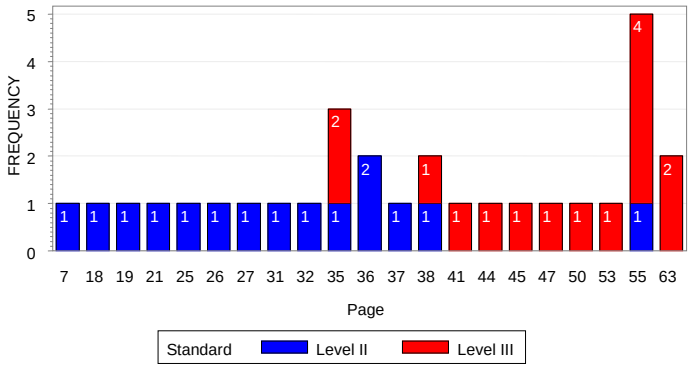

Round 1 Panelist Agreement Data
STAAR Chemistry
FREQUENCY
0 1 2 3 4 5 6
1 1 3 1 1 1 2 1 1 1 1 2 6 1 1 1 1
17 22 23 28 29 34 35 38 41 42 53 54 55 56 60 61 62
Page
Standard
Level II
Level III


Round 2 Panelist Agreement Data
STAAR Chemistry
FREQUENCY
0 1 2 3 4 5 6 7 8 9 10
2 3 1 1 1 1 1 2 1 1 1 10 1
22 23 24 27 29 30 31 35 40 49 50 55 60
Page
Standard
Level II
Level III

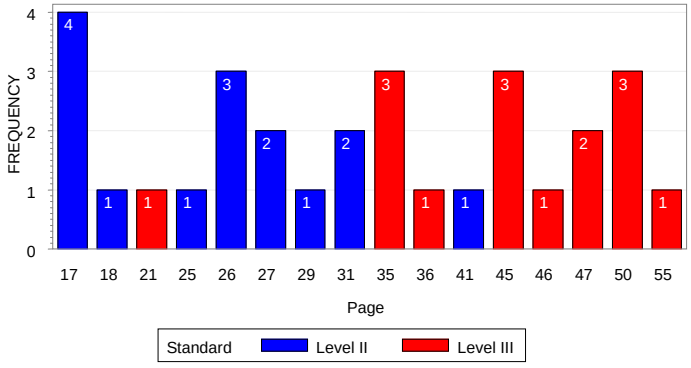

Round 3 Panelist Agreement Data
STAAR Chemistry
FREQUENCY
0 1 2 3 4 5 6 7 8 9 10 11
1 2 2 1 2 1 1 1 1 1 1 11 1
22 23 24 27 28 30 35 36 38 40 53 55 60
Page
Standard
Level II
Level III

PHYSICS

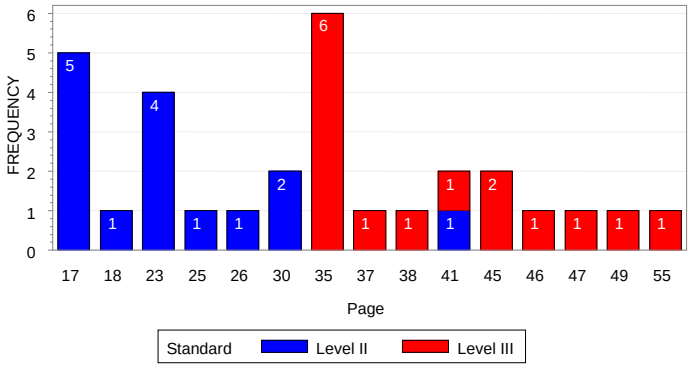
Round 1 Panelist Agreement Data
STAAR Physics



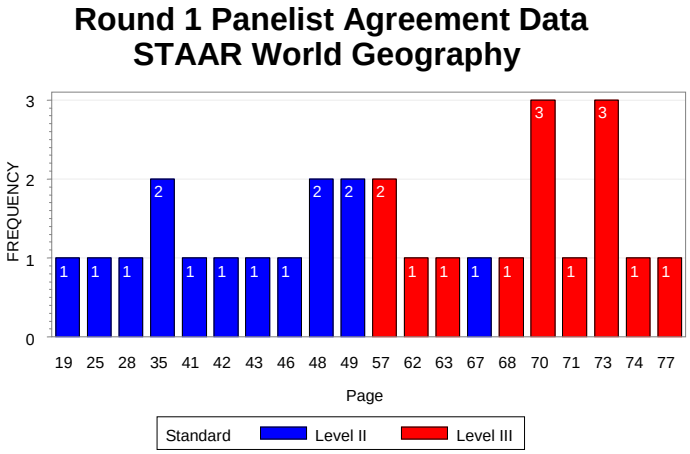
Round 2 Panelist Agreement Data
STAAR Physics

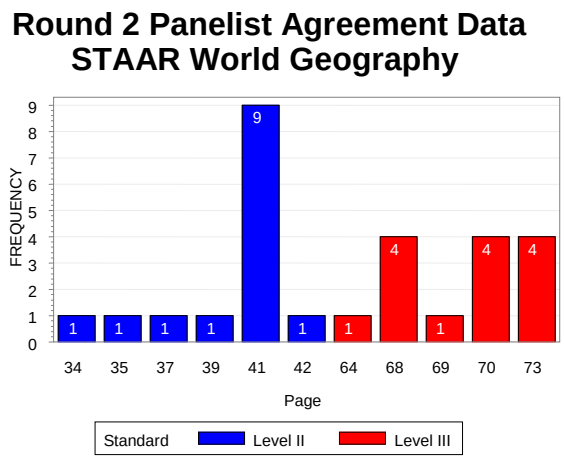


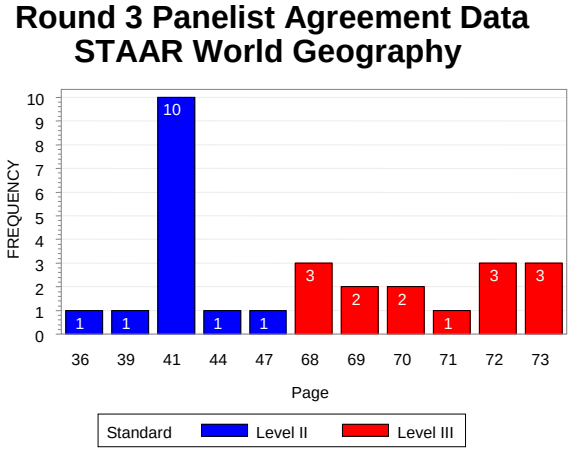
Round 3 Panelist Agreement Data
STAAR Physics



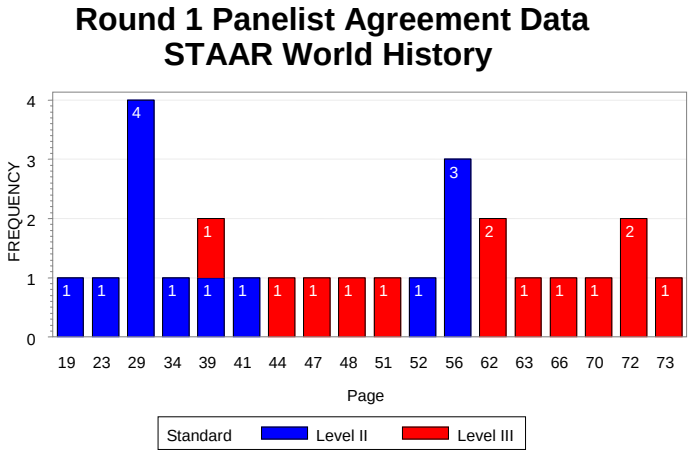
WORLD GEOGRAPHY

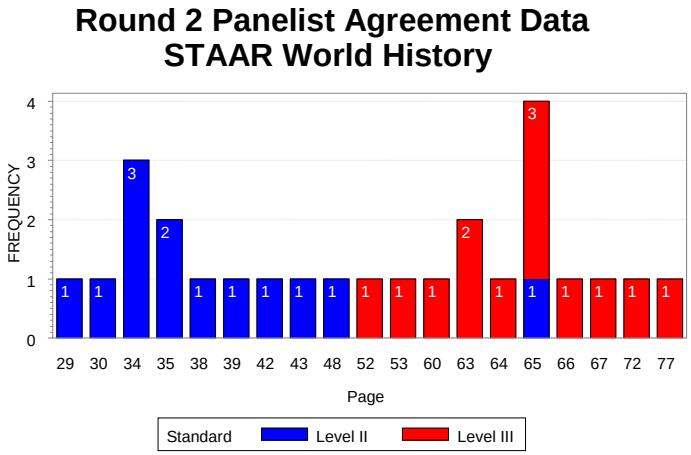


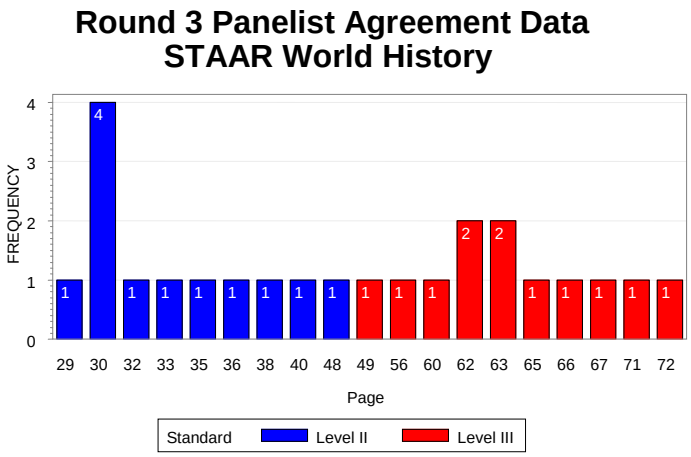




WORLD HISTORY

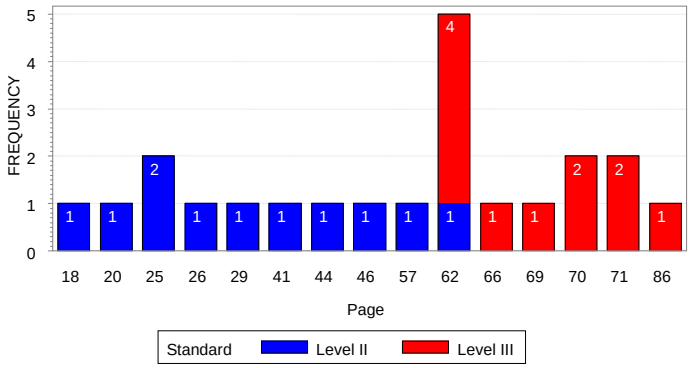




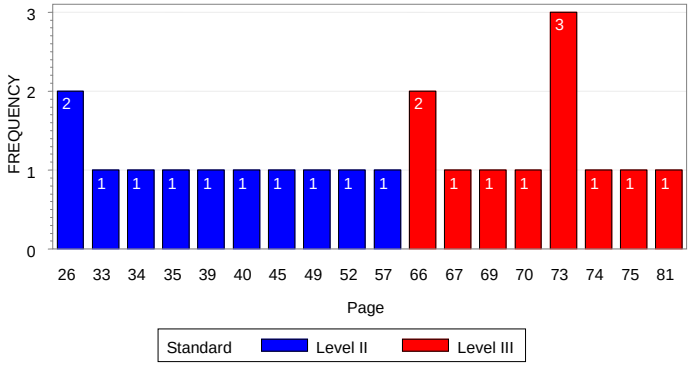


U.S. HISTORY

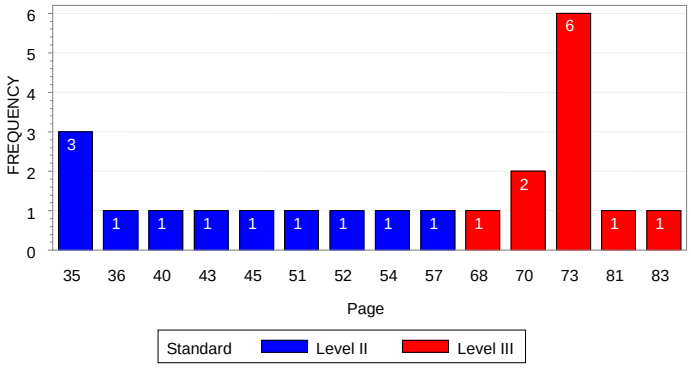
Round 1 Panelist Agreement Data
STAAR U.S. History



Round 2 Panelist Agreement Data
STAAR U.S. History

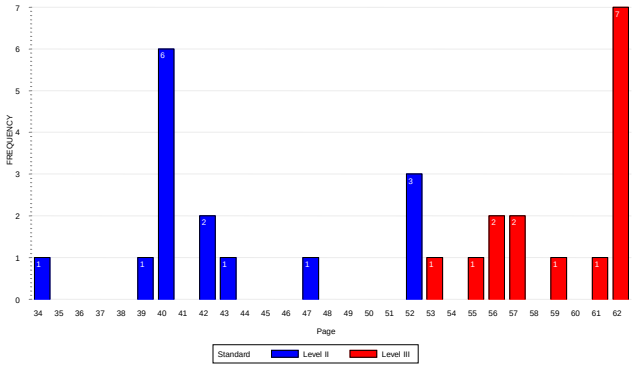


Round 3 Panelist Agreement Data
STAAR U.S. History

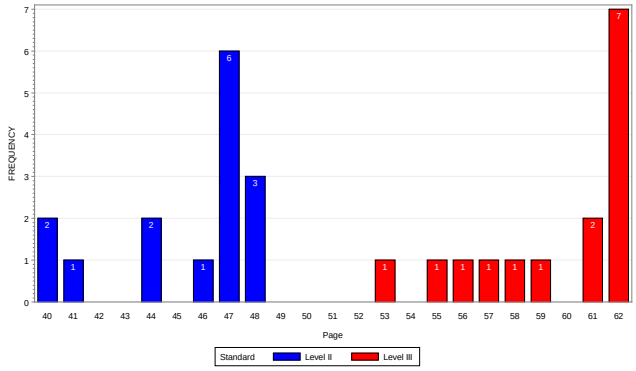


GRADE 3 MATHEMATICS

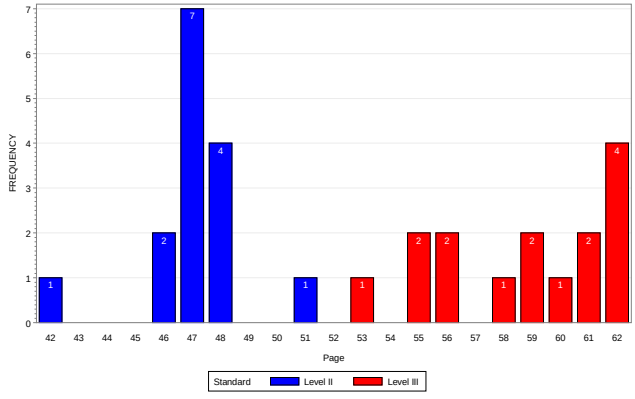
Round 1 Panelist Agreement Data
STAAR Grade 03 Mathematics



Round 2 Panelist Agreement Data
STAAR Grade 03 English Mathematics

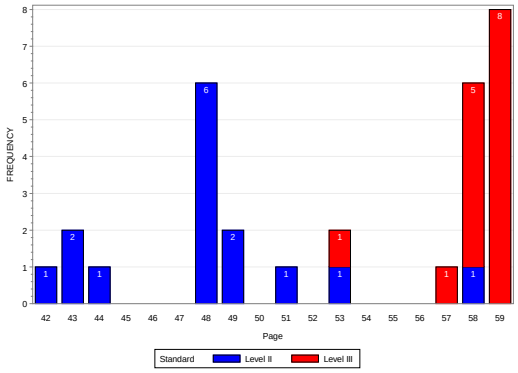


Round 3 Panelist Agreement Data
STAAR Grade 03 English Mathematics

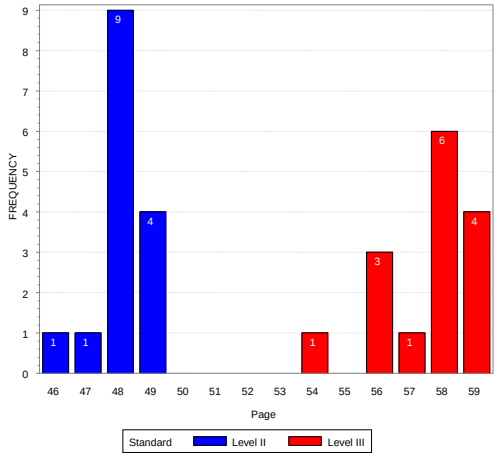


STAAR GRADE 4 MATHEMATICS

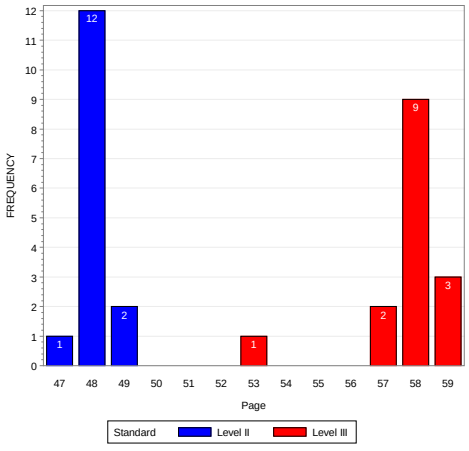
Round 1 Panelist Agreement Data
STAAR Grade 04 English Mathematics



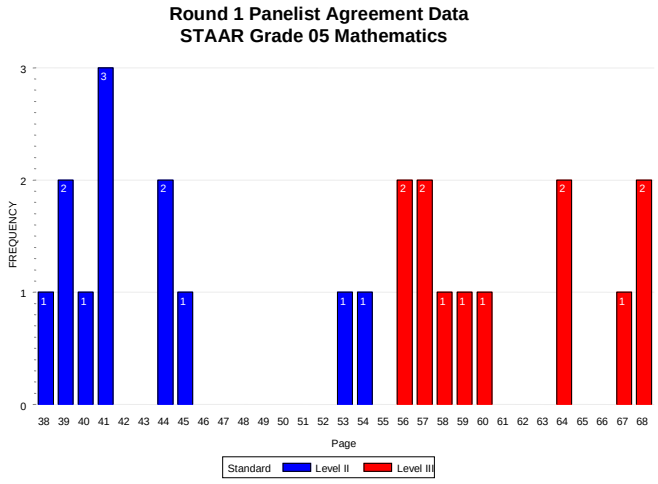
Round 2 Panelist Agreement Data
STAAR Grade 04 English Mathematics

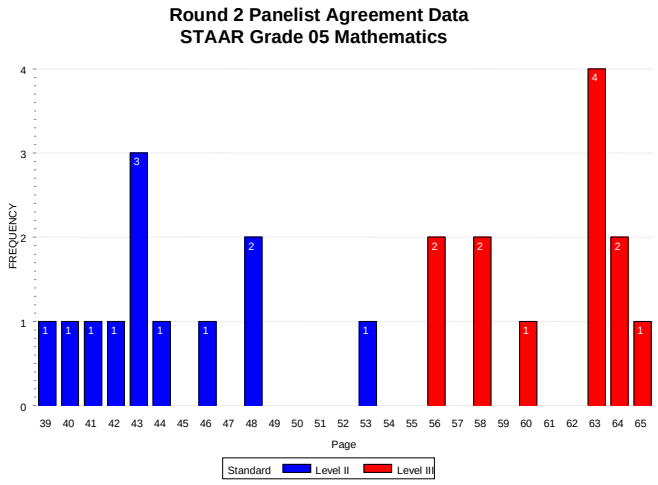


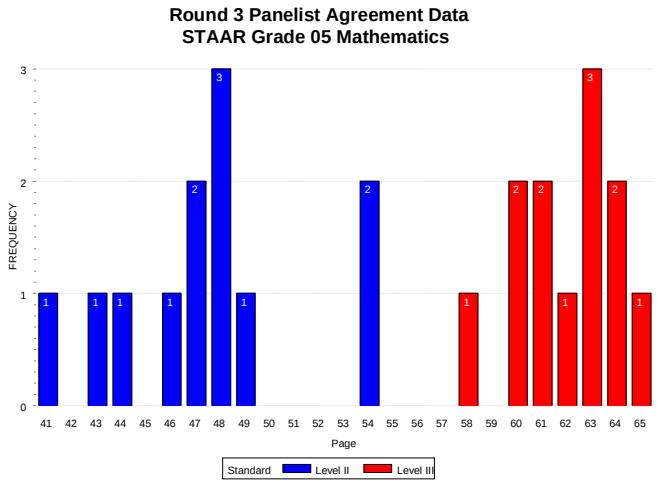
Round 3 Panelist Agreement Data
STAAR Grade 04 English Mathematics



STAAR GRADE 5 MATHEMATICS

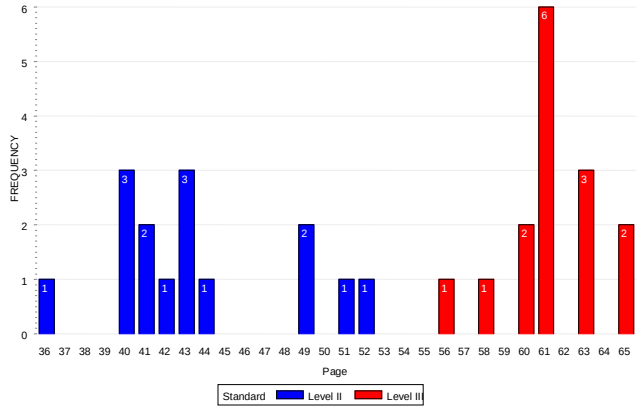




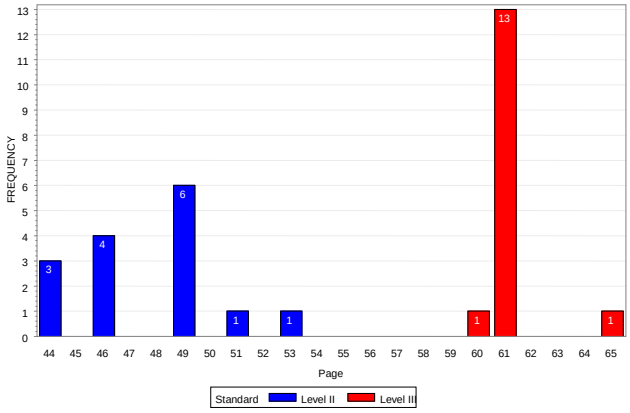


GRADE 6 MATHEMATICS

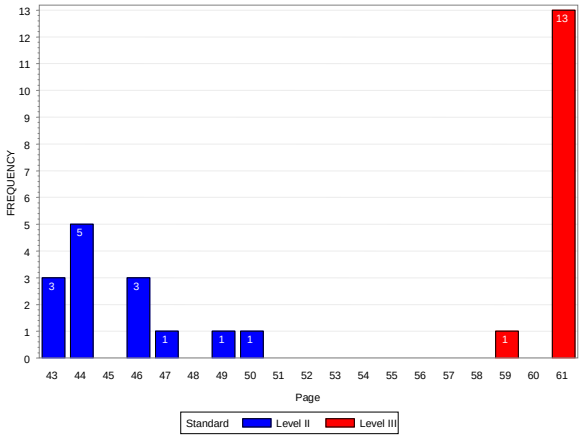
Round 1 Panelist Agreement Data
STAAR Grade 06 Mathematics



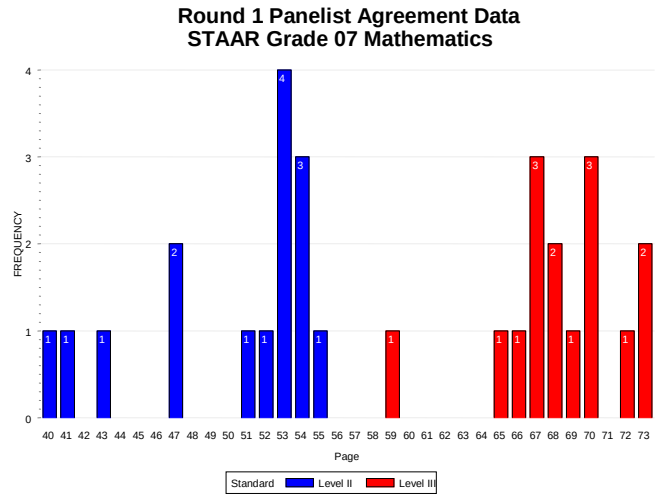
Round 2 Panelist Agreement Data
STAAR Grade 06 Mathematics

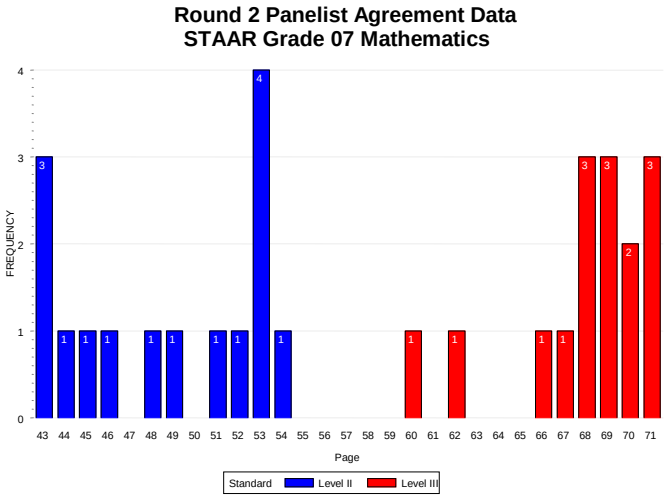


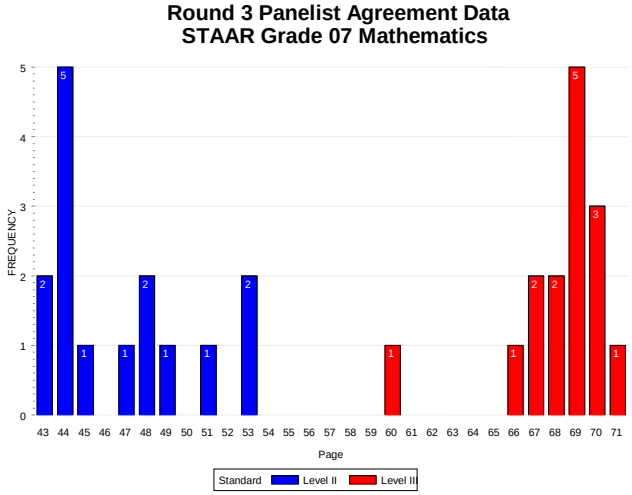
Round 3 Panelist Agreement Data
STAAR Grade 06 Mathematics



GRADE 7 MATHEMATICS

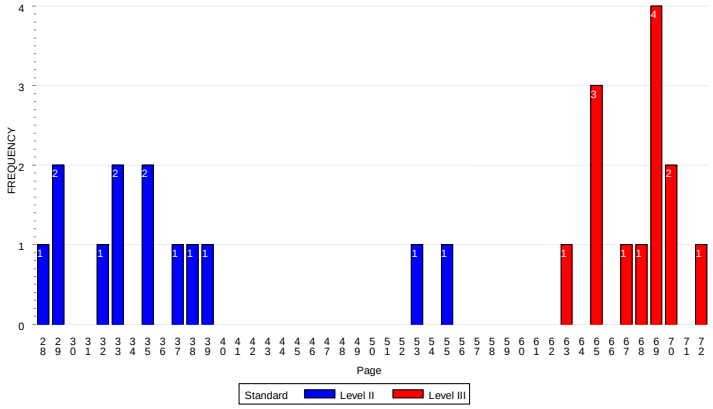






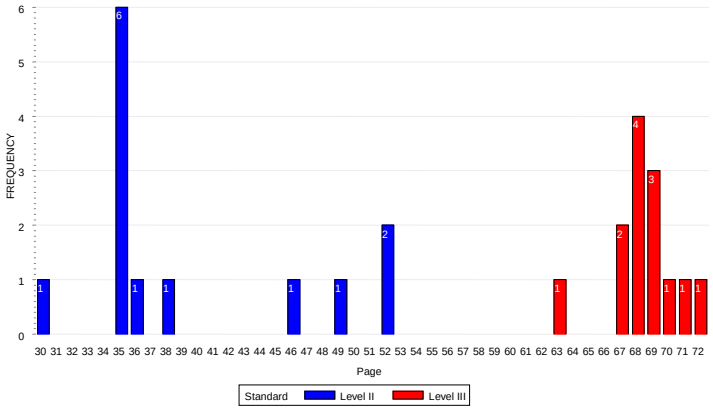
GRADE 8 MATHEMATICS

Round 1 Panelist Agreement Data
STAAR Grade 08 Mathematics

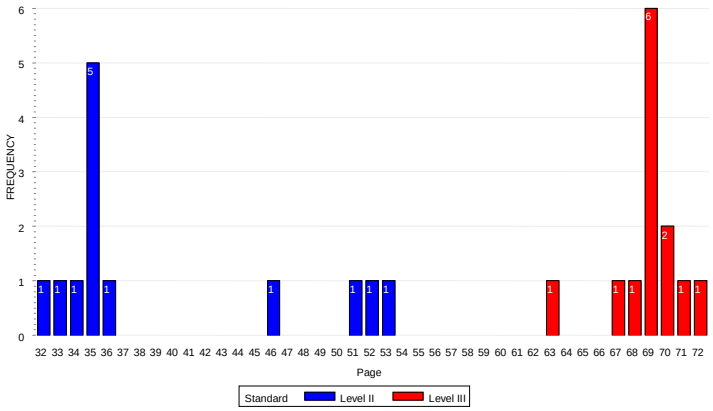


* Numbers along the horizontal axis are stacked due to the large range of panelists' judgments.

Round 2 Panelist Agreement Data
STAAR Grade 08 Mathematics

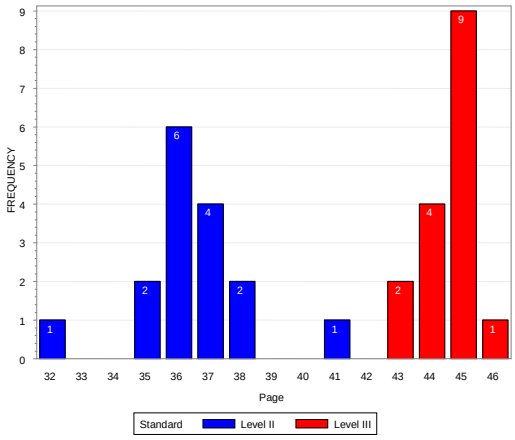


Round 3 Panelist Agreement Data
STAAR Grade 08 Mathematics

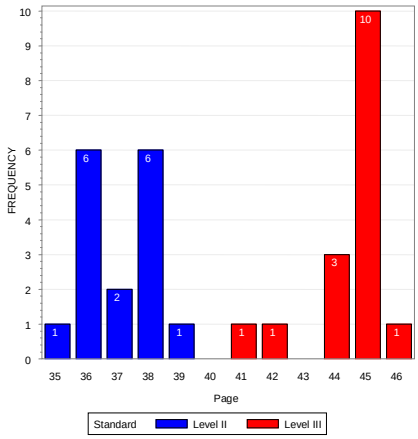


GRADE 3 ENGLISH READING

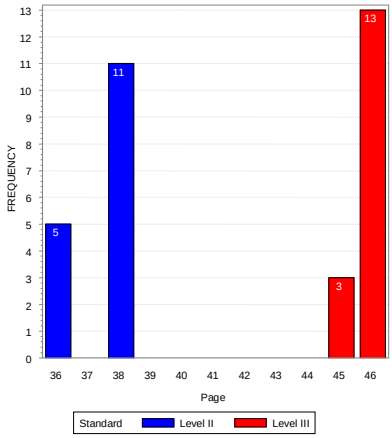
Round 1 Panelist Agreement Data
STAAR Grade 03 English Reading



Round 2 Panelist Agreement Data
STAAR Grade 03 English Reading

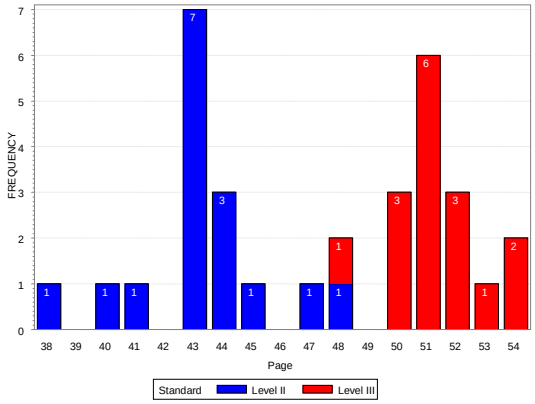


Round 3 Panelist Agreement Data
STAAR Grade 03 English Reading

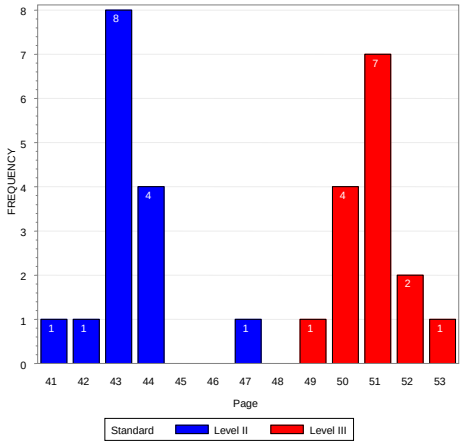


GRADE 4 ENGLISH READING

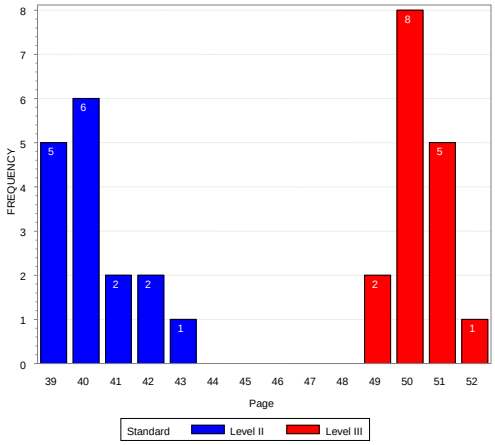
Round 1 Panelist Agreement Data
STAAR Grade 04 English Reading



Round 2 Panelist Agreement Data
STAAR Grade 04 English Reading

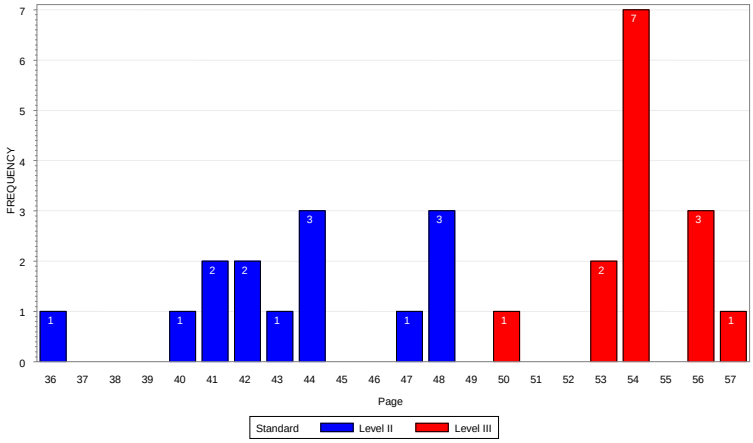


Round 3 Panelist Agreement Data
STAAR Grade 04 English Reading

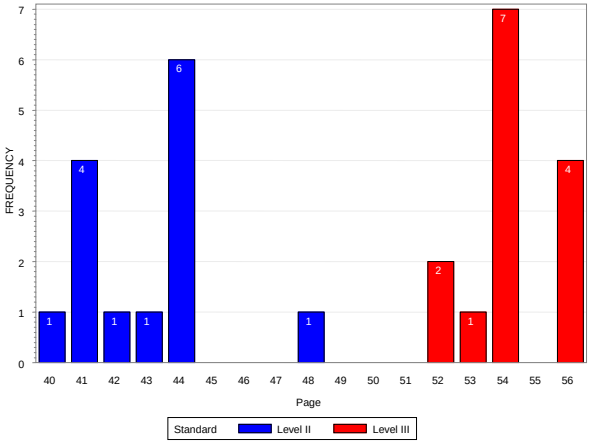


GRADE 5 ENGLISH READING

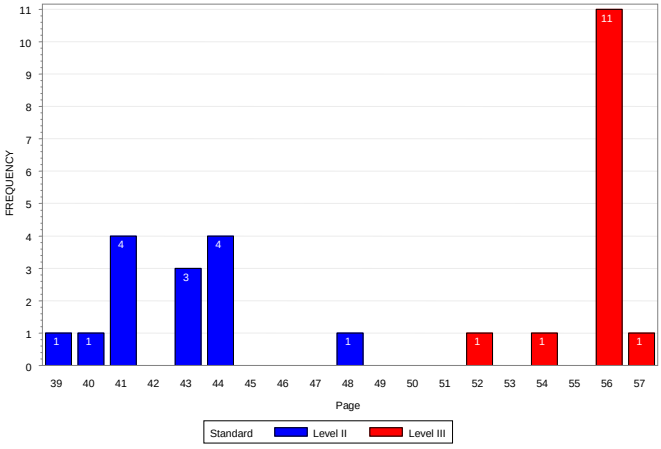
Round 1 Panelist Agreement Data
STAAR Grade 05 English Reading



Round 2 Panelist Agreement Data
STAAR Grade 05 English Reading

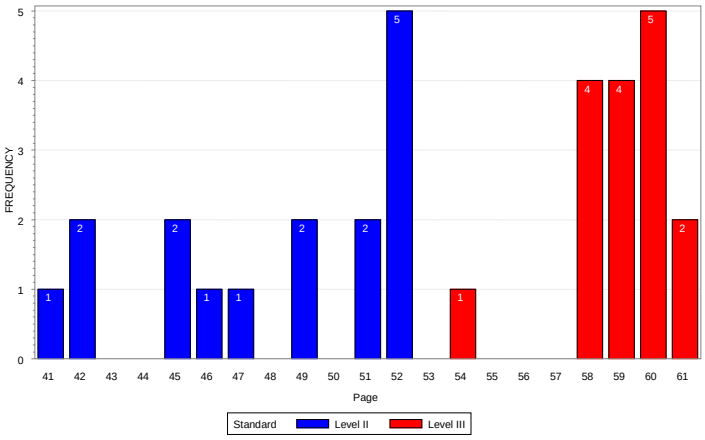


Round 3 Panelist Agreement Data
STAAR Grade 05 English Reading

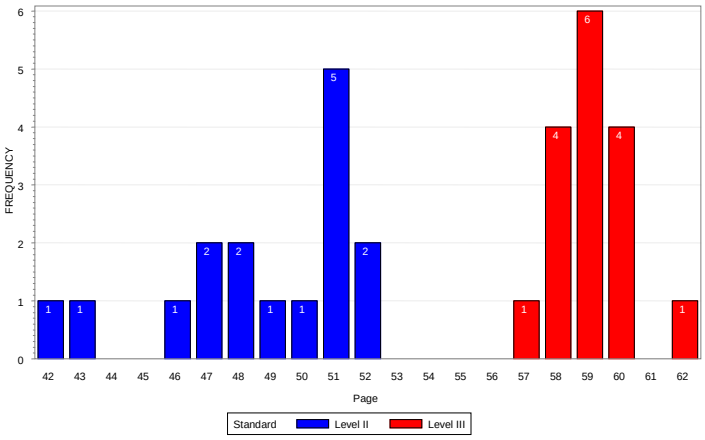


GRADE 6 READING

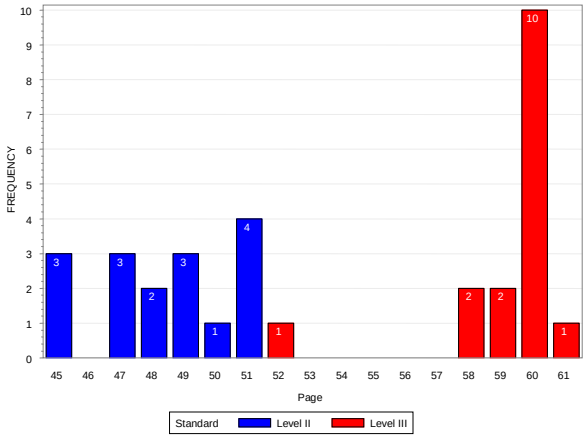
Round 1 Panelist Agreement Data
STAAR Grade 06 Reading



Round 2 Panelist Agreement Data
STAAR Grade 06 Reading

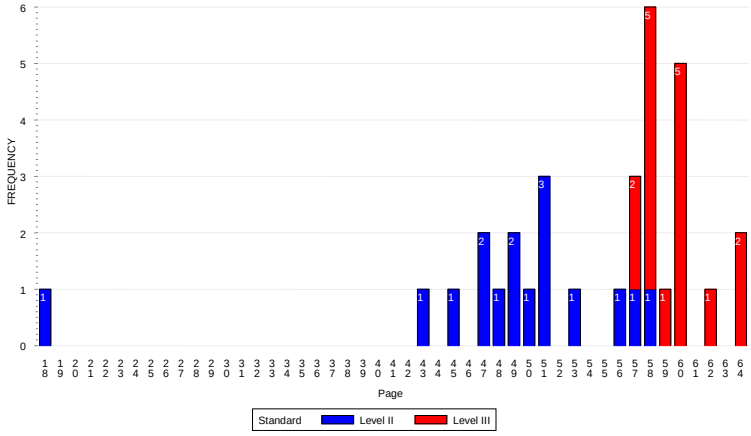


Round 3 Panelist Agreement Data
STAAR Grade 06 Reading



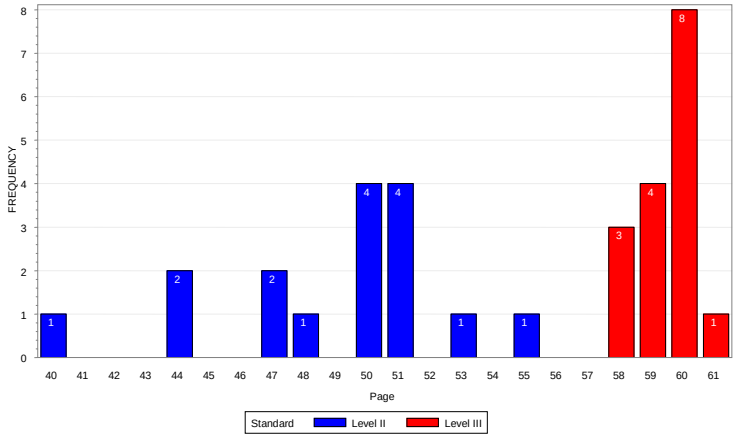
GRADE 7 READING

Round 1 Panelist Agreement Data
STAAR Grade 07 Reading

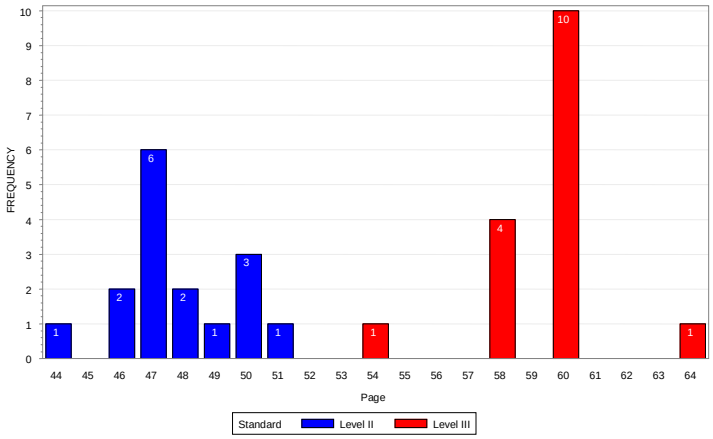


* Numbers along the horizontal axis are stacked due to the large range of panelists' judgments.

Round 2 Panelist Agreement Data
STAAR Grade 07 Reading

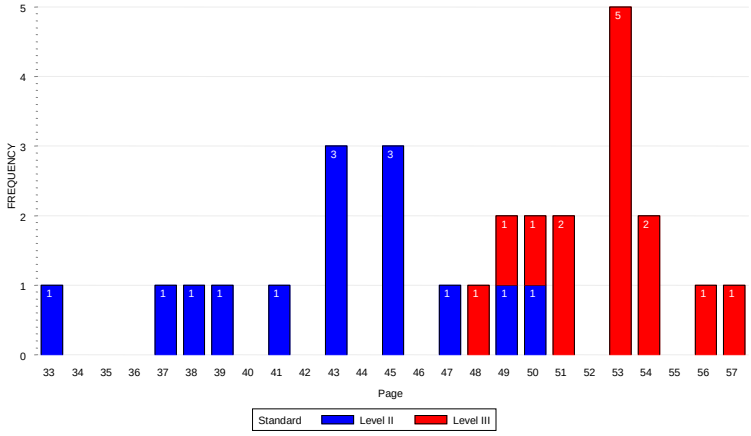


Round 3 Panelist Agreement Data
STAAR Grade 07 Reading

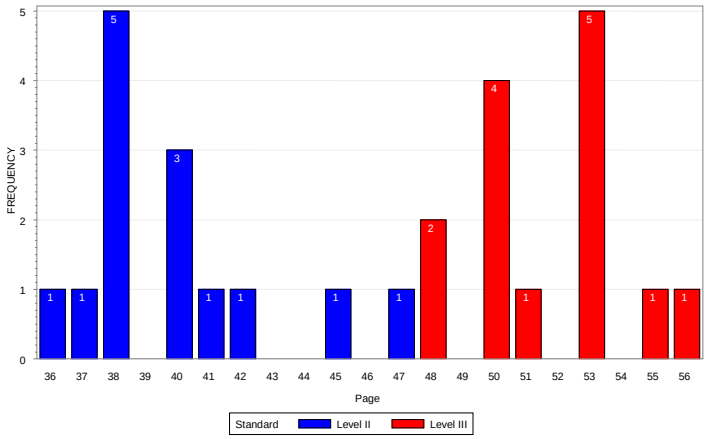


GRADE 8 READING

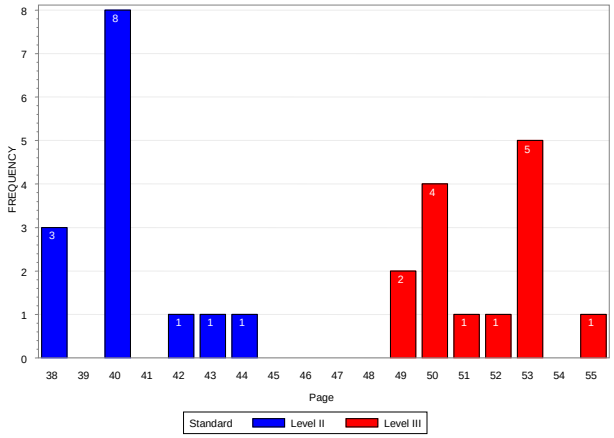
Round 1 Panelist Agreement Data
STAAR Grade 08 Reading



Round 2 Panelist Agreement Data
STAAR Grade 08 Reading

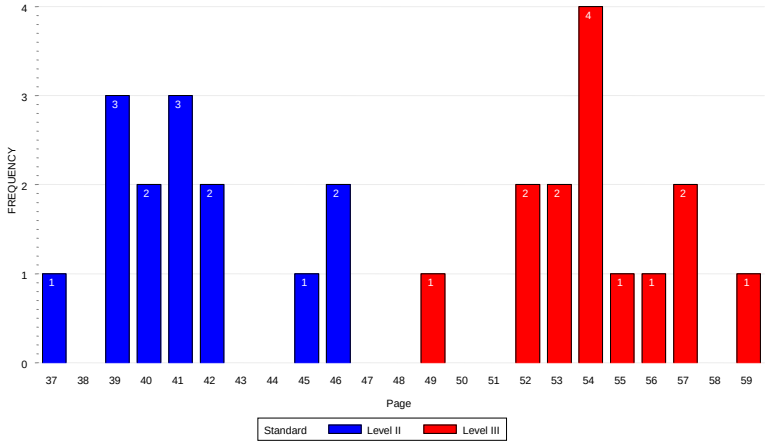


Round 3 Panelist Agreement Data
STAAR Grade 08 Reading

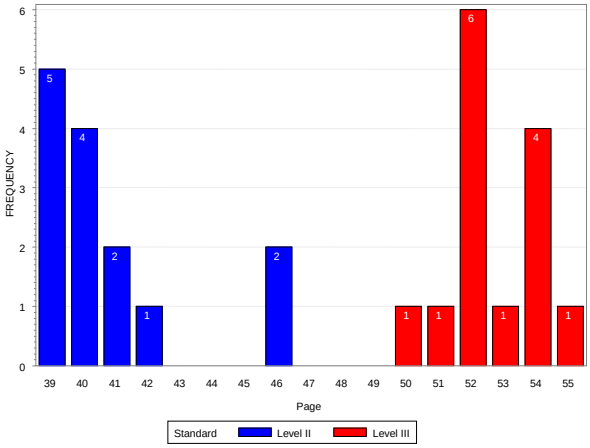


GRADE 3 SPANISH READING

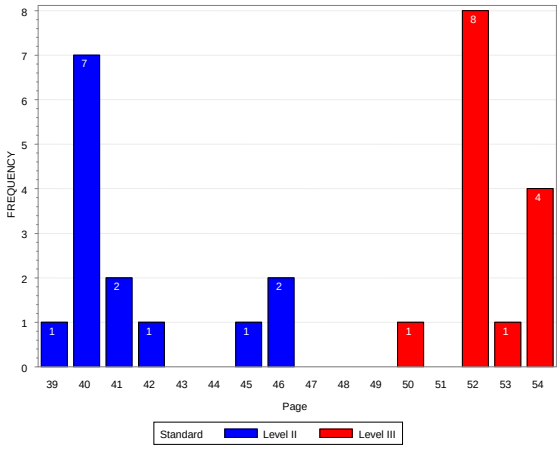
Round 1 Panelist Agreement Data
STAAR Grade 03 Spanish Reading



Round 2 Panelist Agreement Data
STAAR Grade 03 Spanish Reading

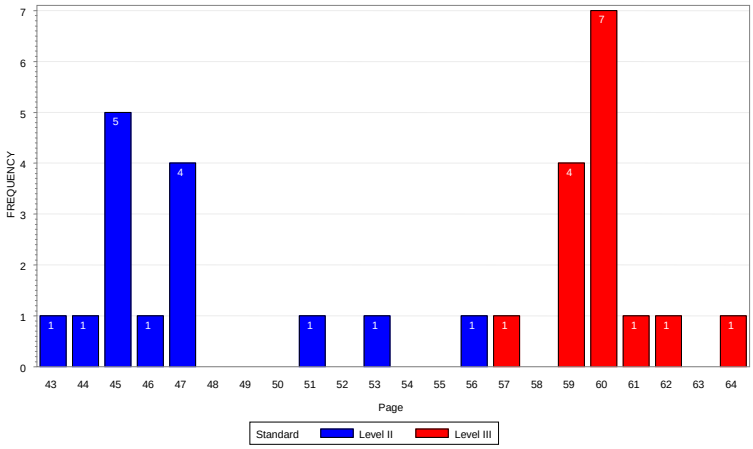


Round 3 Panelist Agreement Data
STAAR Grade 03 Spanish Reading

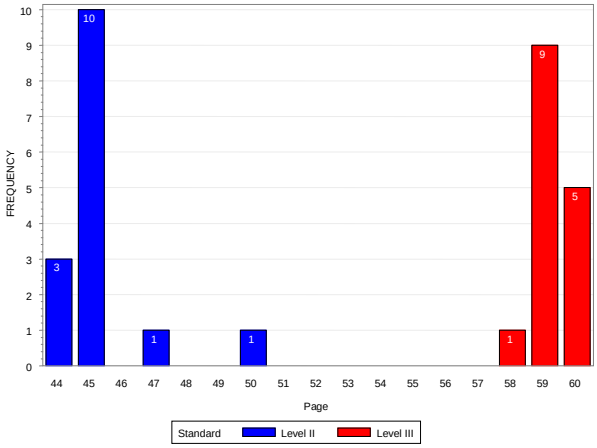


GRADE 4 SPANISH READING

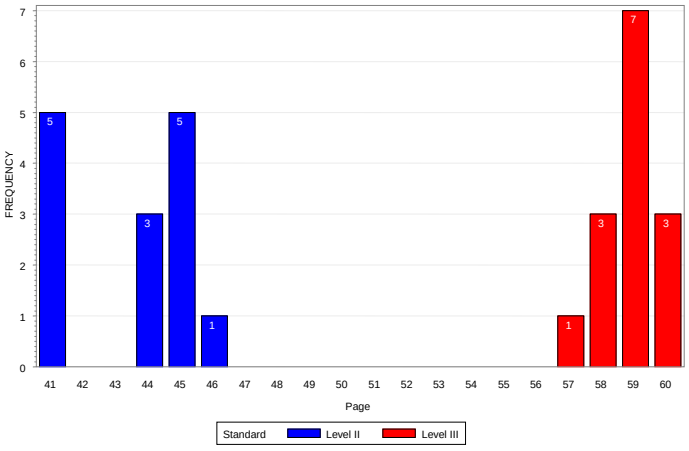
Round 1 Panelist Agreement Data
STAAR Grade 04 Spanish Reading



Round 2 Panelist Agreement Data
STAAR Grade 04 Spanish Reading

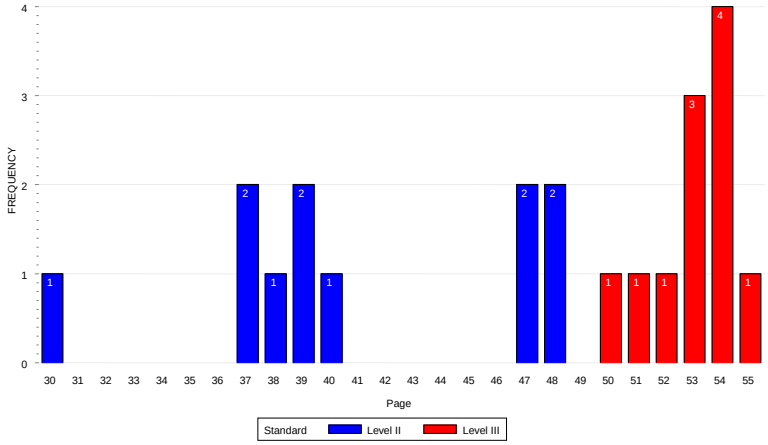


Round 3 Panelist Agreement Data
STAAR Grade 04 Spanish Reading

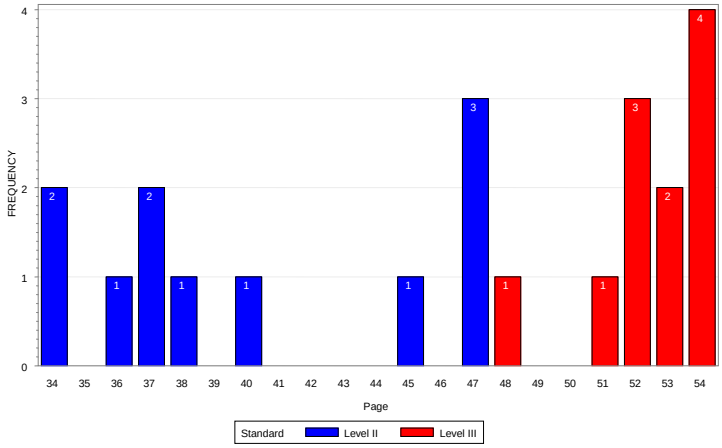


GRADE 5 SPANISH READING

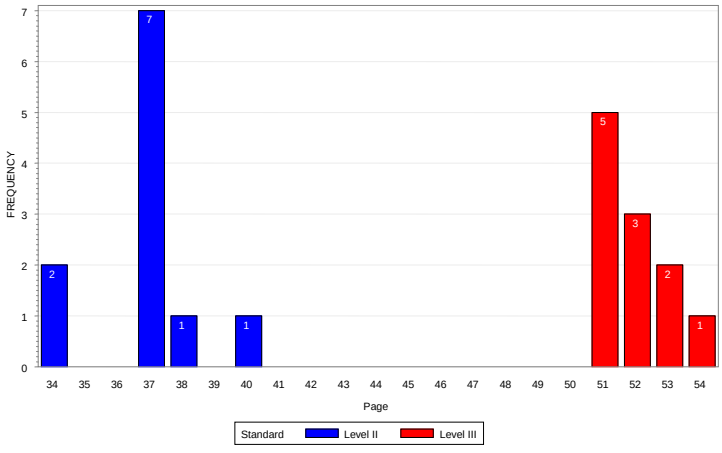
Round 1 Panelist Agreement Data
STAAR Grade 05 Spanish Reading



Round 2 Panelist Agreement Data
STAAR Grade 05 Spanish Reading

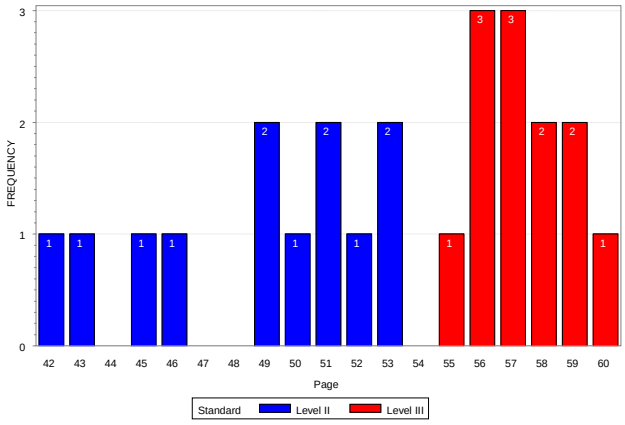


Round 3 Panelist Agreement Data
STAAR Grade 05 Spanish Reading

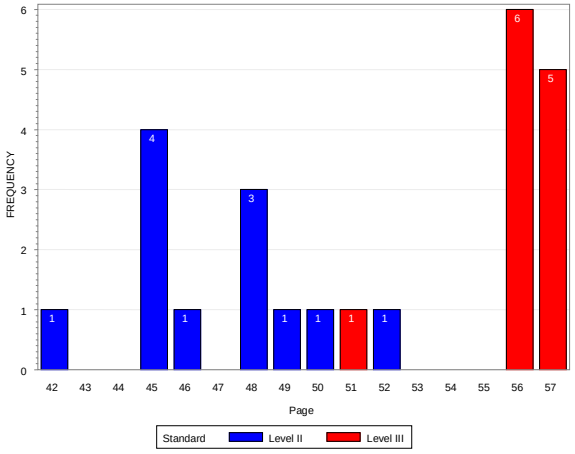


GRADE 5 SCIENCE

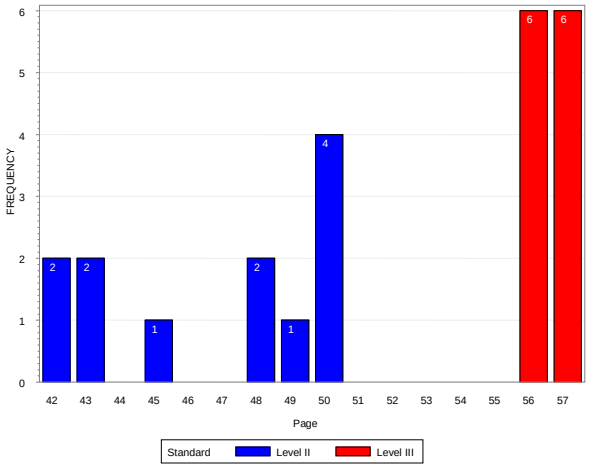
Round 1 Panelist Agreement Data
STAAR Grade 05 Science



Round 2 Panelist Agreement Data
STAAR Grade 05 Science

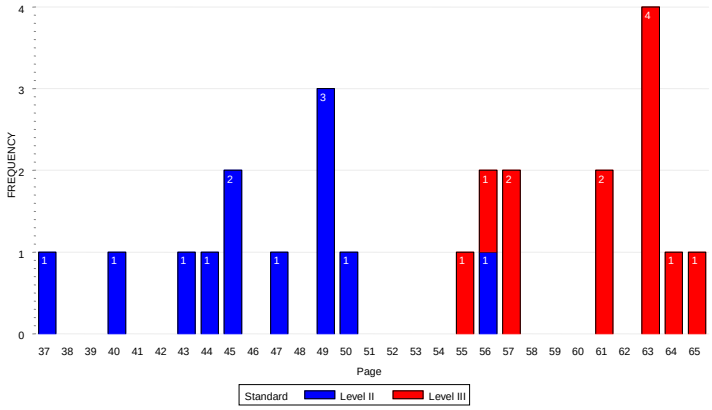


Round 3 Panelist Agreement Data
STAAR Grade 05 Science

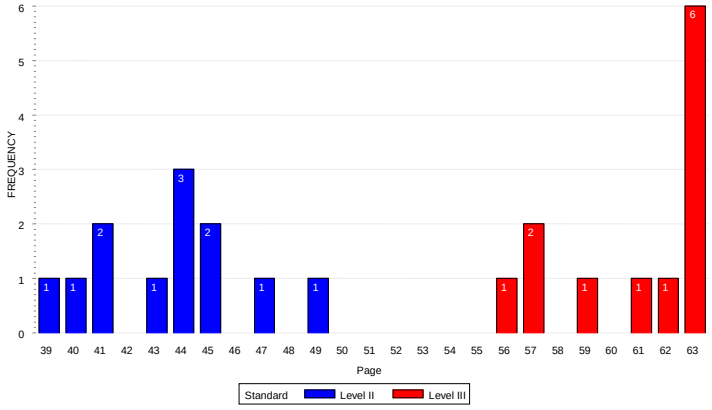


GRADE 8 SCIENCE

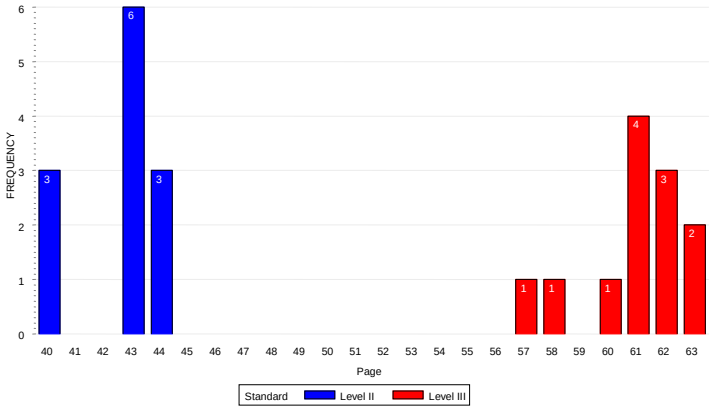
Round 1 Panelist Agreement Data
STAAR Grade 08 Science



Round 2 Panelist Agreement Data
STAAR Grade 08 Science

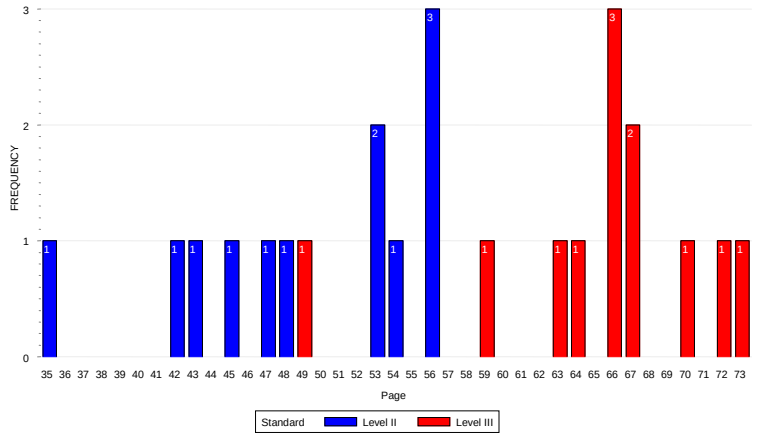


Round 3 Panelist Agreement Data
STAAR Grade 08 Science

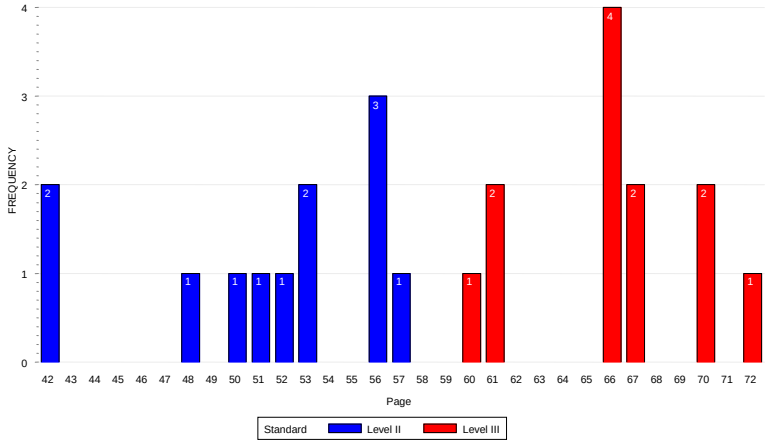


GRADE 8 SOCIAL STUDIES

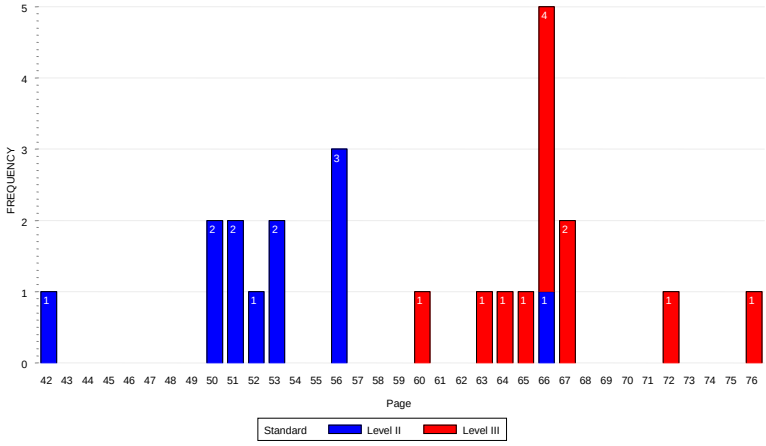
Round 1 Panelist Agreement Data
STAAR Grade 08 Social Studies



Round 2 Panelist Agreement Data
STAAR Grade 08 Social Studies

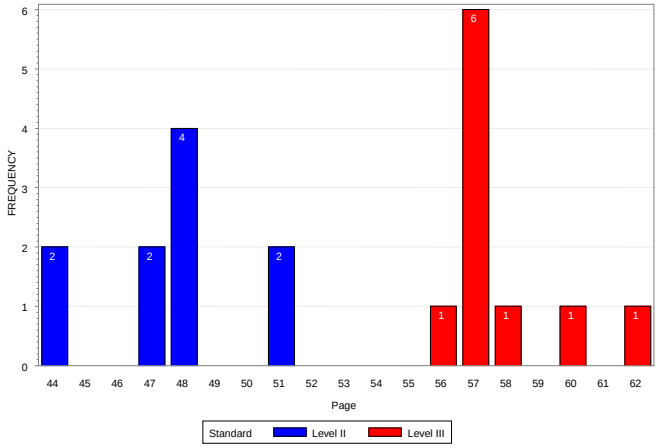


Round 3 Panelist Agreement Data
STAAR Grade 08 Social Studies

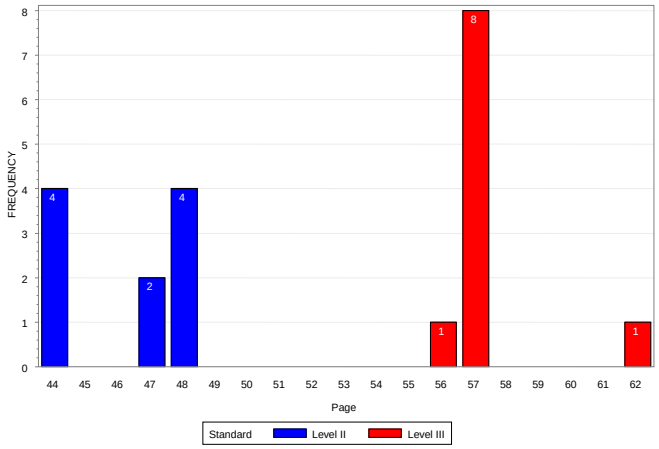


GRADE 4 ENGLISH WRITING

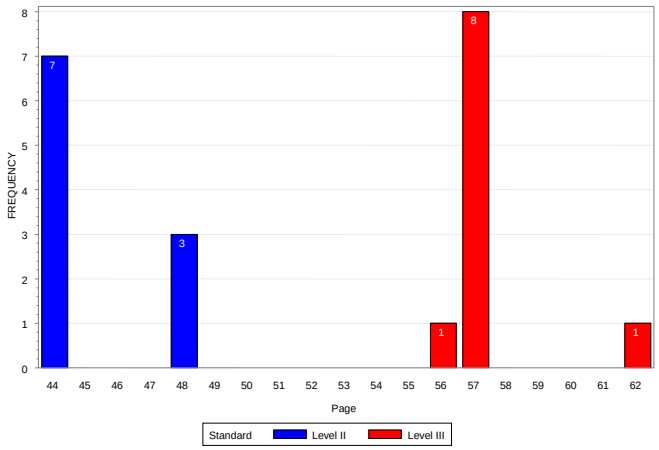
Round 1 Panelist Agreement Data
STAAR Grade 04 English Writing



Round 2 Panelist Agreement Data
STAAR Grade 04 English Writing

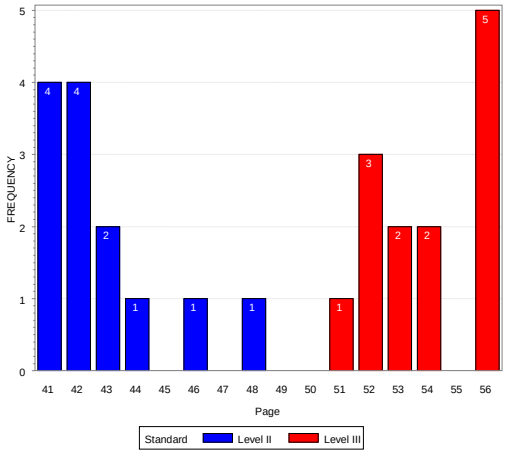


Round 3 Panelist Agreement Data
STAAR Grade 04 English Writing

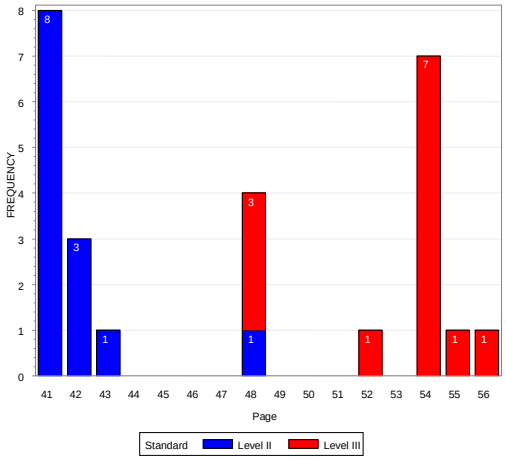


GRADE 7 WRITING

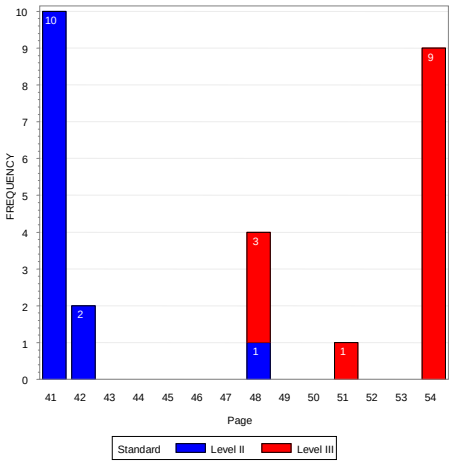
Round 1 Panelist Agreement Data
STAAR Grade 07 Writing



Round 2 Panelist Agreement Data
STAAR Grade 07 Writing

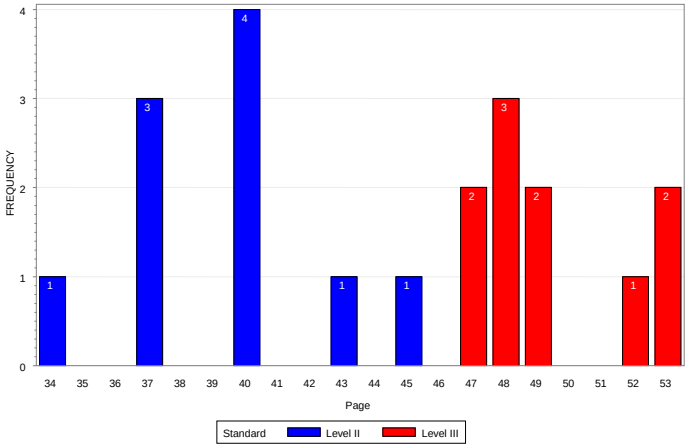


Round 3 Panelist Agreement Data
STAAR Grade 07 Writing

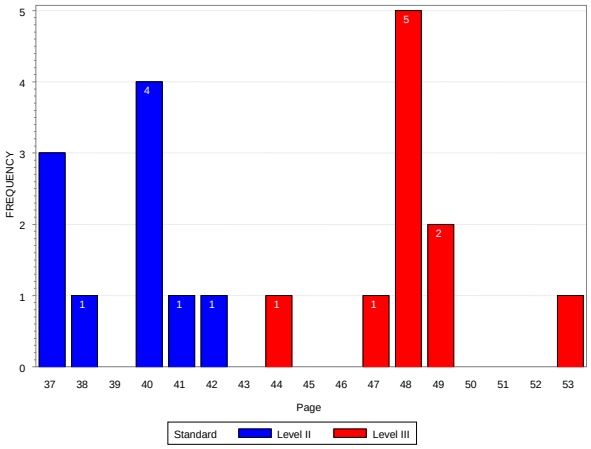


GRADE 4 SPANISH WRITING

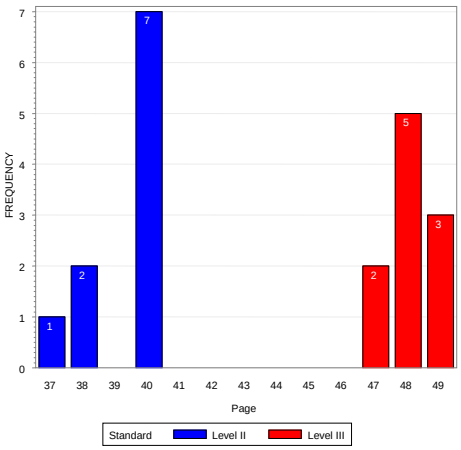
Round 1 Panelist Agreement Data
STAAR Grade 04 Spanish Writing



Round 2 Panelist Agreement Data
STAAR Grade 04 Spanish Writing



Round 3 Panelist Agreement Data
STAAR Grade 04 Spanish Writing



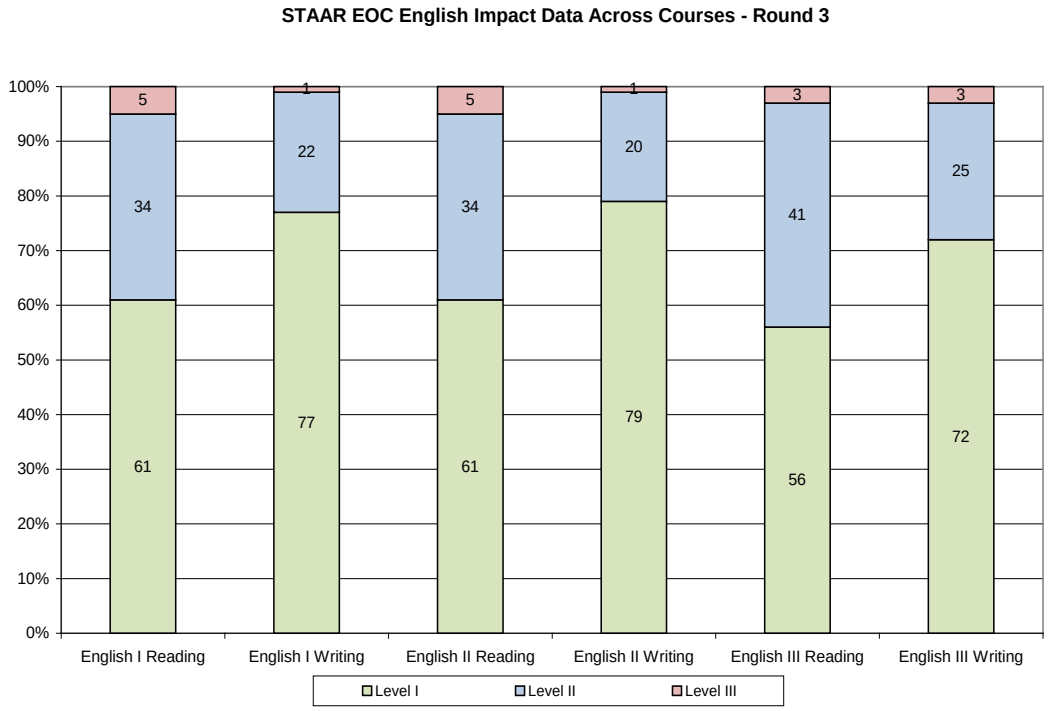
Appendix 19: Estimated Impact Data

This appendix provides the estimated impact data (percentage of students at each performance level) based on the cut score recommendations after Round 3 of the standard-setting committee meetings, cross-course articulation, and the reasonableness review. For the STAAR EOC assessments, the impact data are computed based on student performance on the assessments administered in spring 2011. Refer to Chapter 3 for information about the potential effect of motivation on the impact data from spring 2011. For the STAAR 3–8 assessments, the impact data are computed based on student performance on the spring 2012 administration.

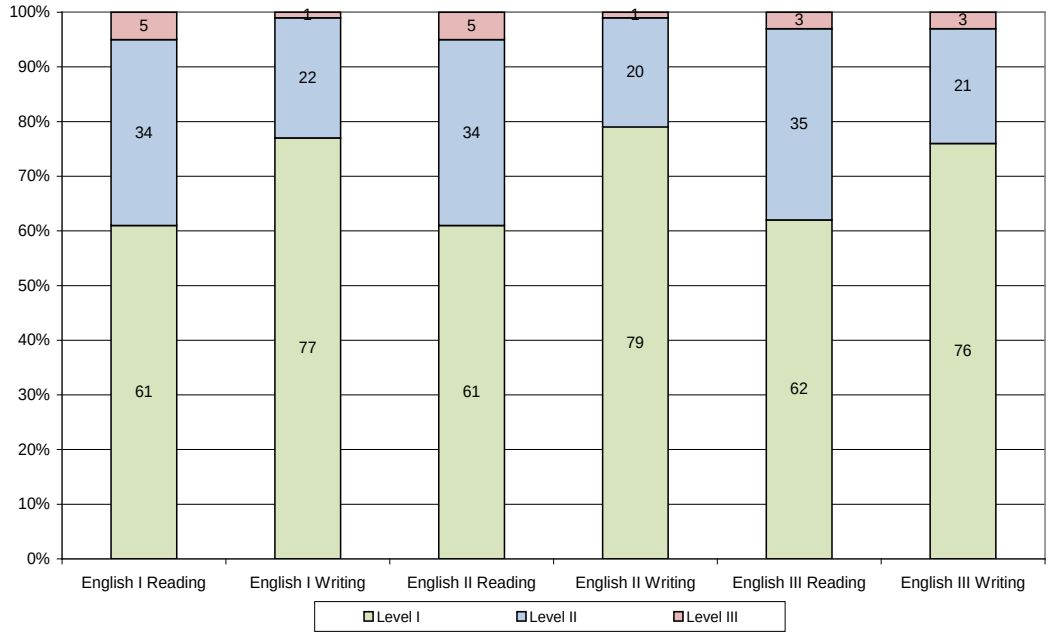
Impact data are shown separately for each content area.

STAAR EOC Assessments

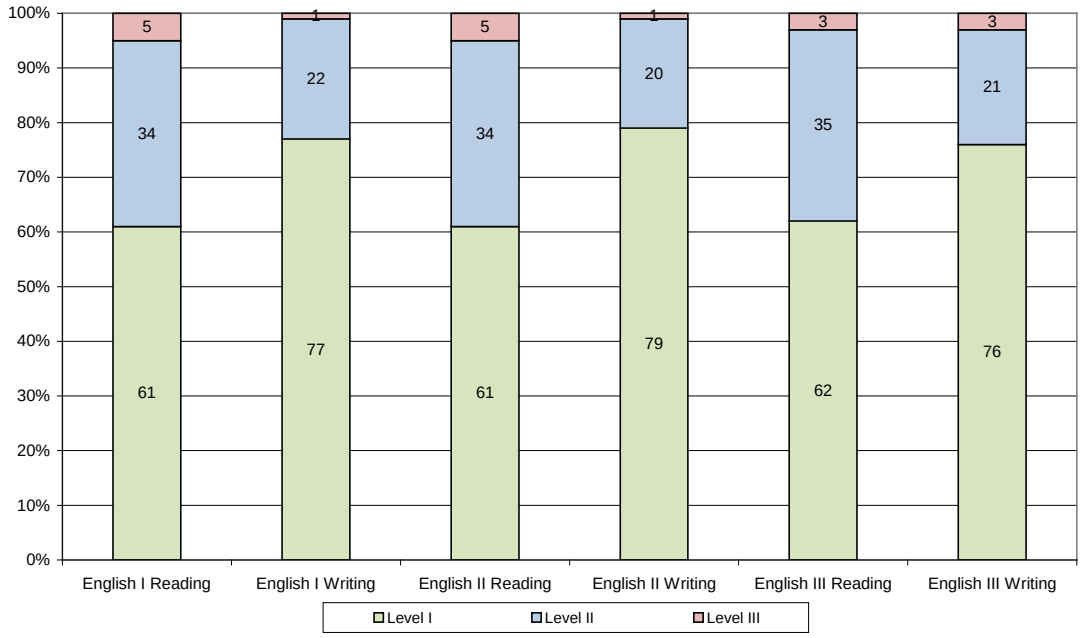
ENGLISH



STAAR EOC English Impact Data Across Courses - Articulation

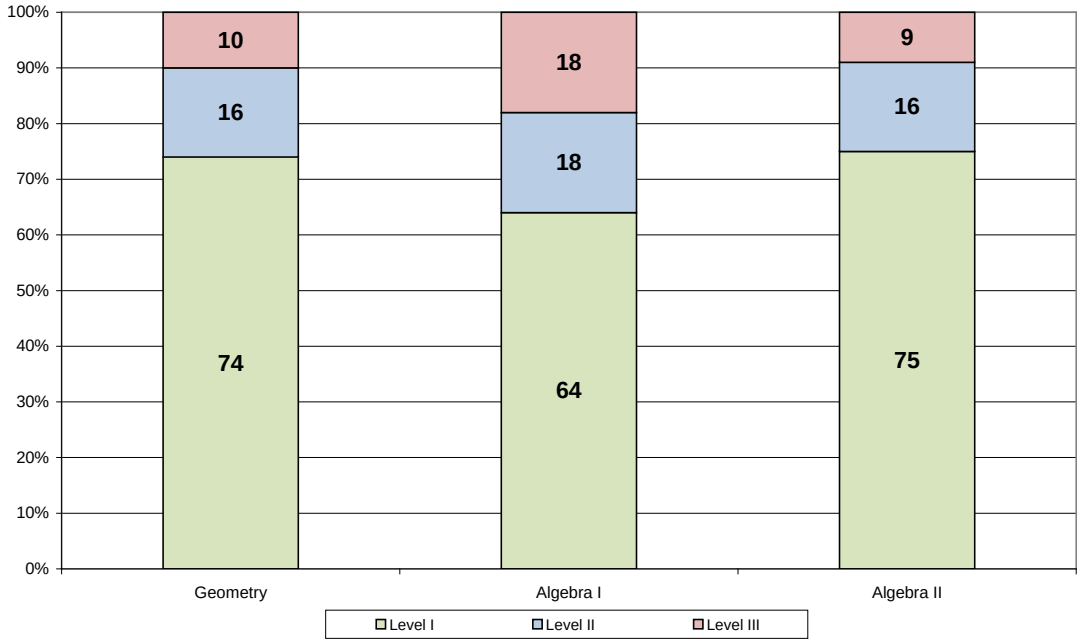


STAAR EOC English Impact Data Across Courses - Reasonableness Review

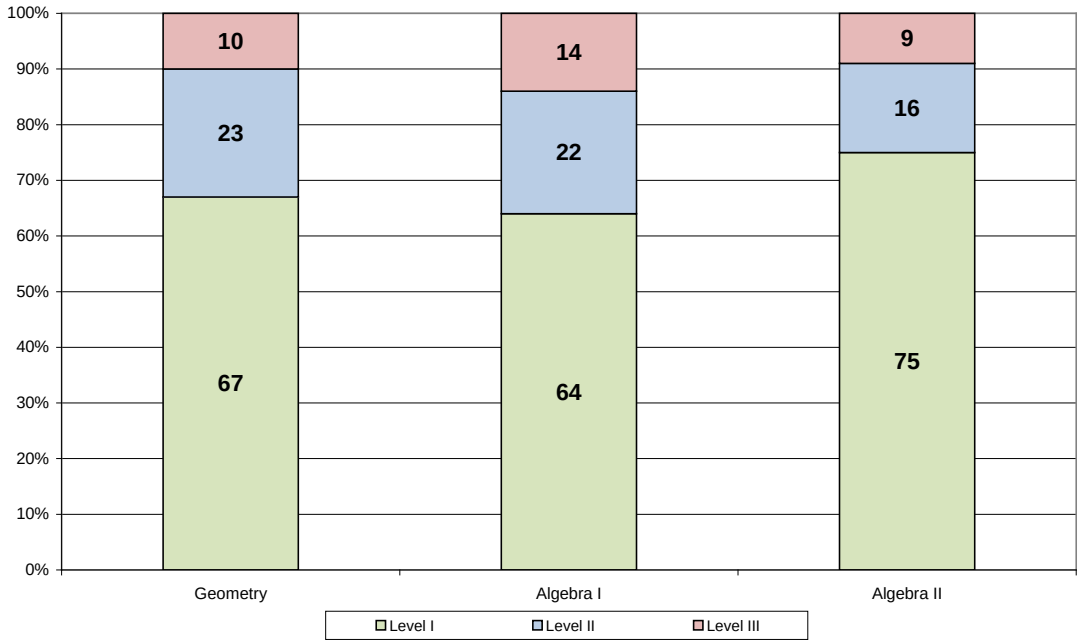


MATHEMATICS

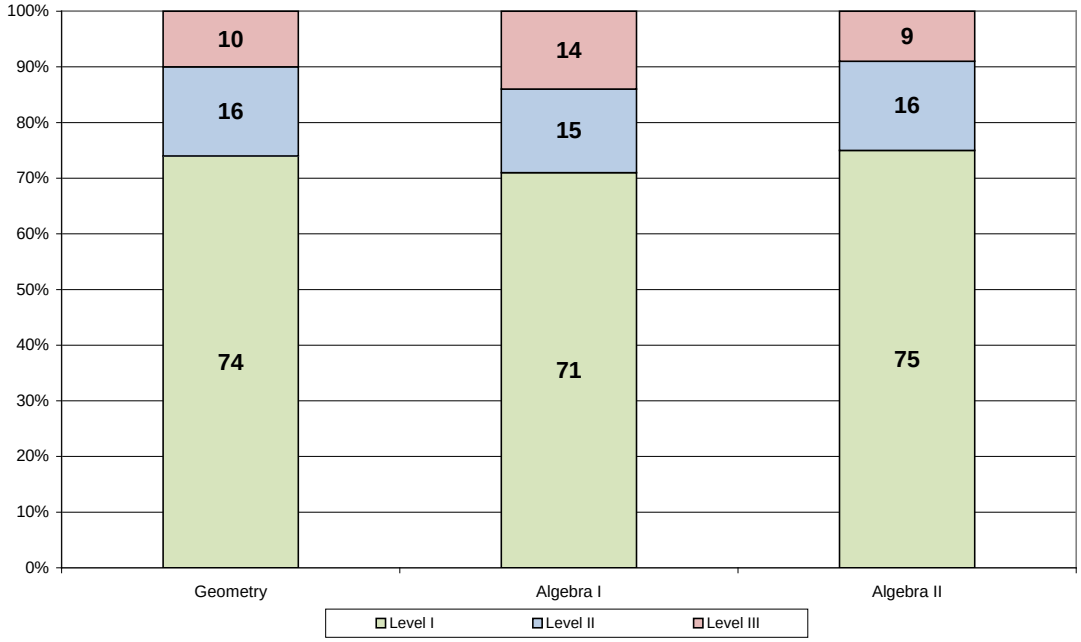
STAAR EOC Mathematics Impact Data Across Courses - Round 3



STAAR EOC Mathematics Impact Data Across Courses - Articulation

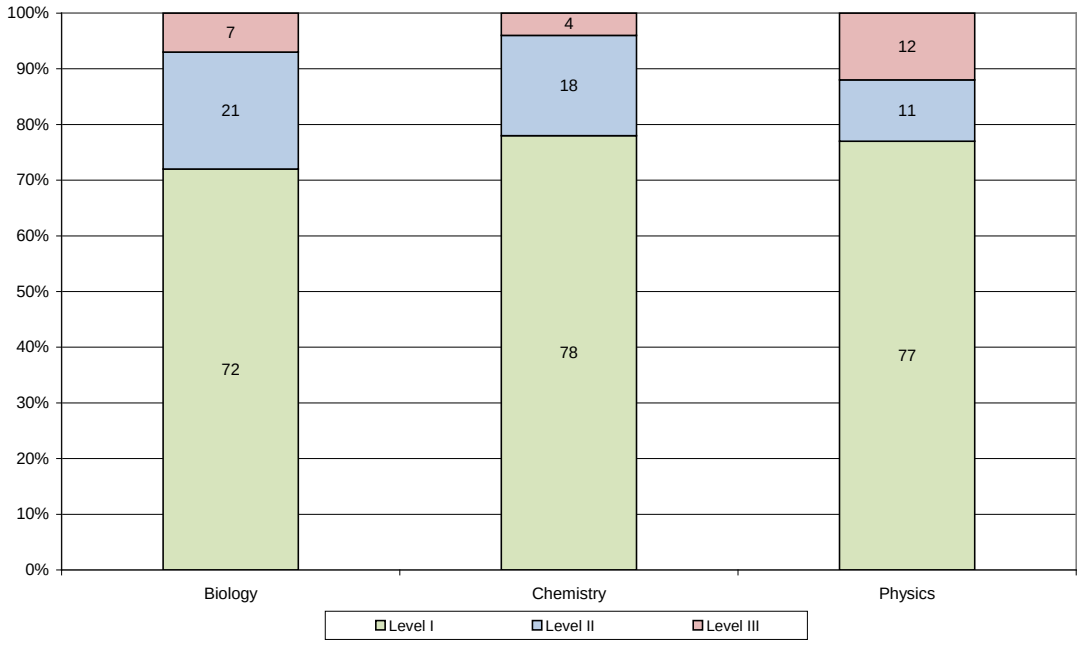


STAAR EOC Mathematics Impact Data Across Courses - Reasonableness Review

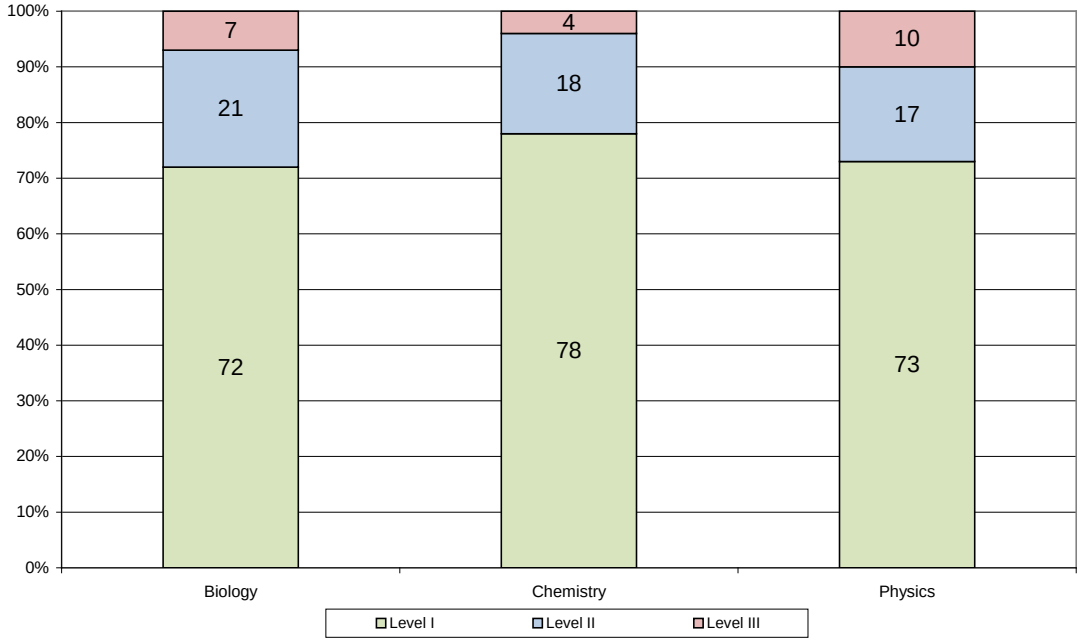


SCIENCE

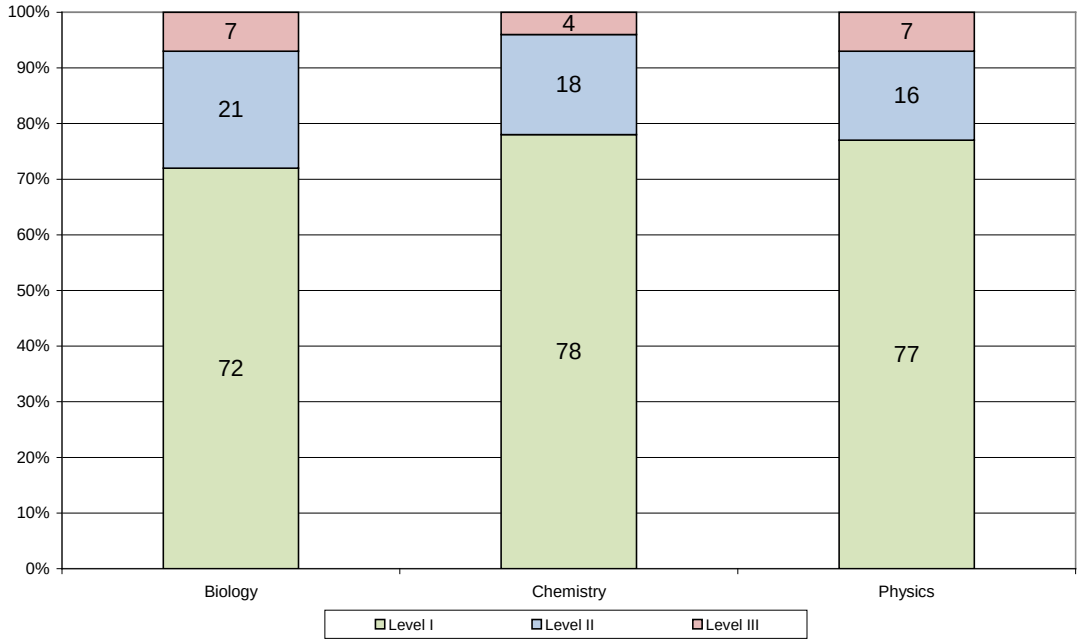
STAAR EOC Science Impact Data Across Courses - Round 3



STAAR EOC Science Impact Data Across Courses - Articulation

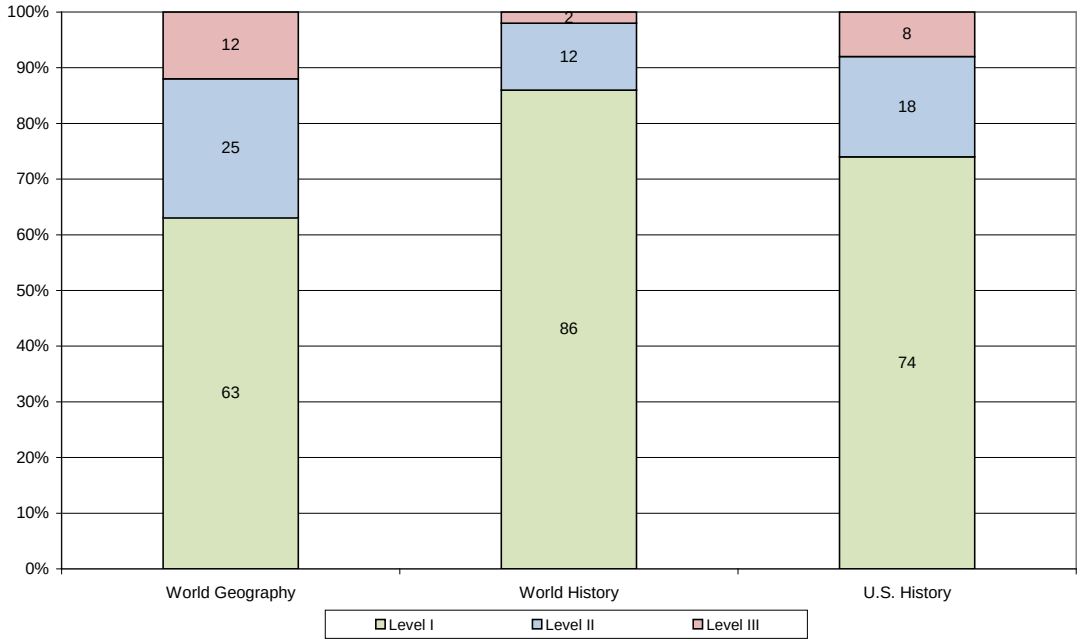


STAAR EOC Science Impact Data Across Courses - Reasonableness Review

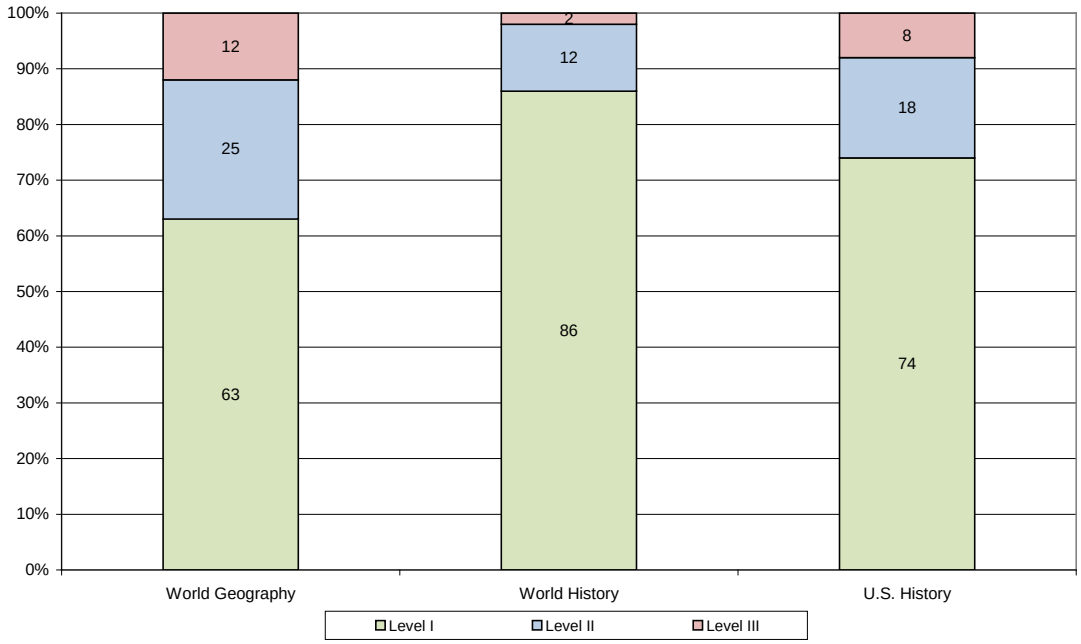


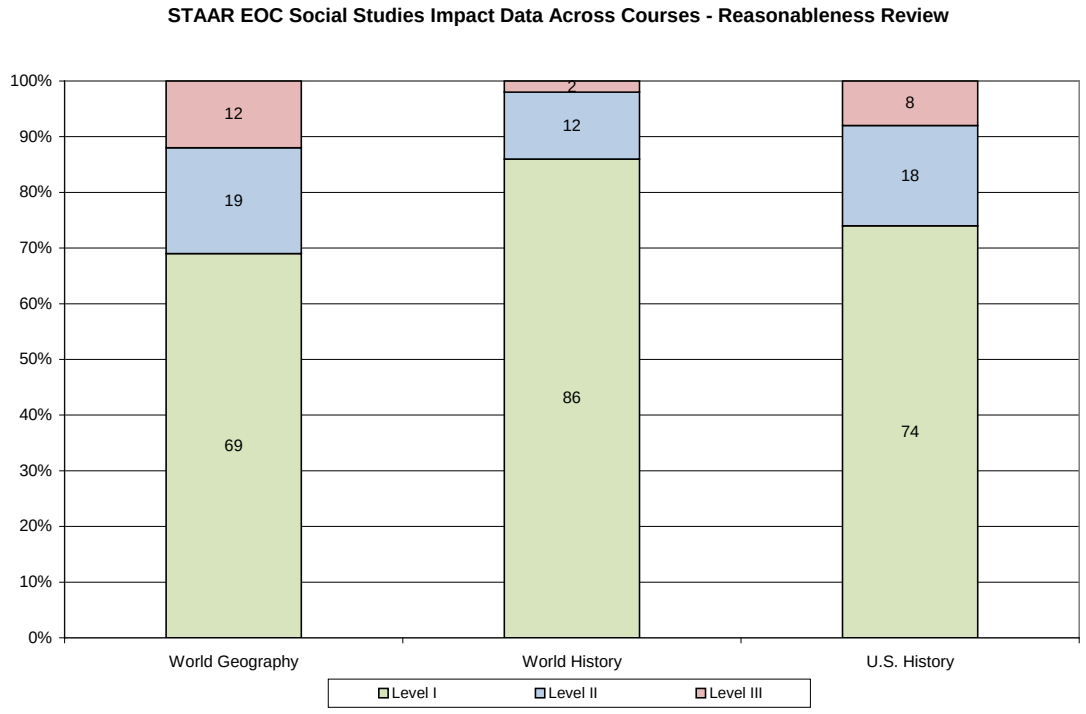
SOCIAL STUDIES

STAAR EOC Social Studies Impact Data Across Courses - Round 3



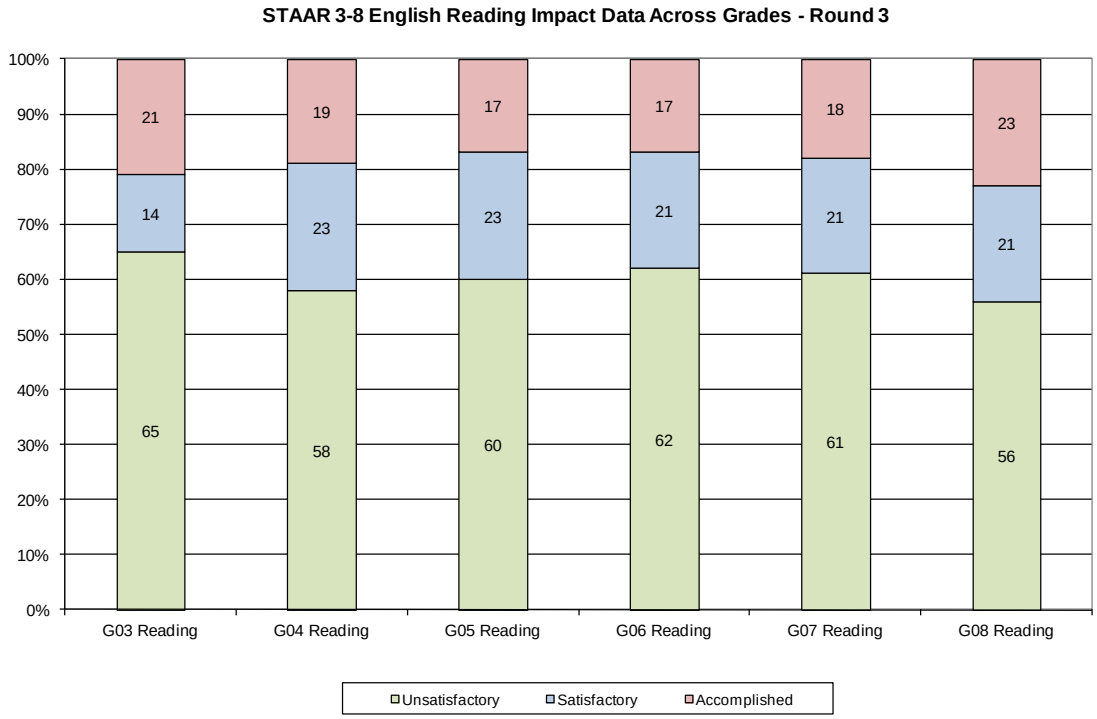
STAAR EOC Social Studies Impact Data Across Courses - Articulation



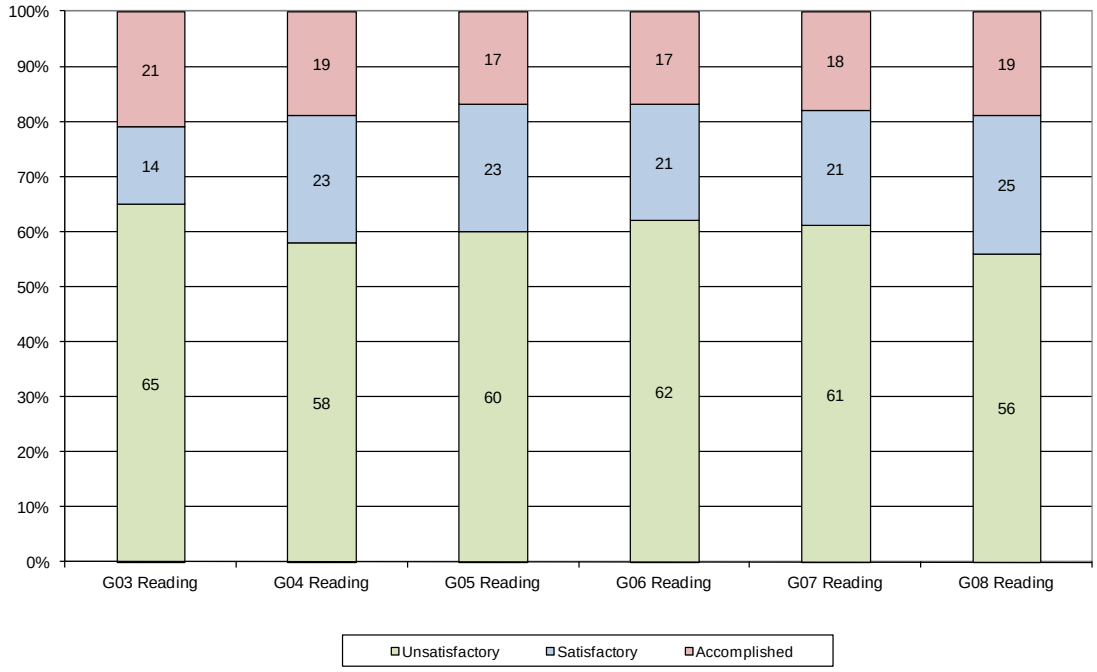


STAAR 3–8 Assessments

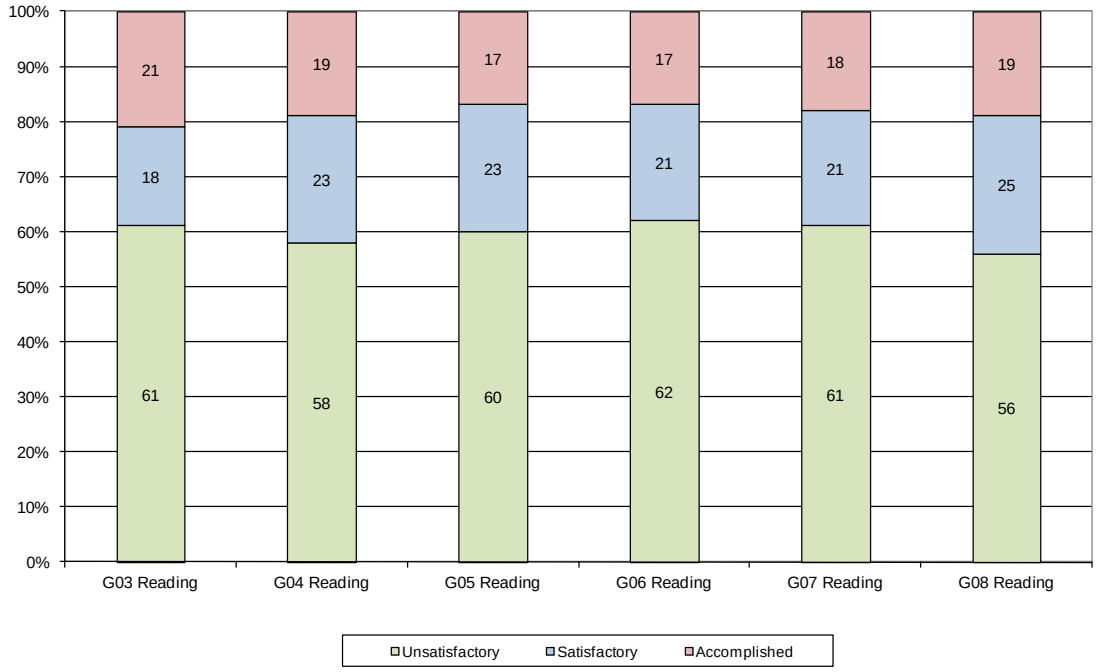
ENGLISH READING



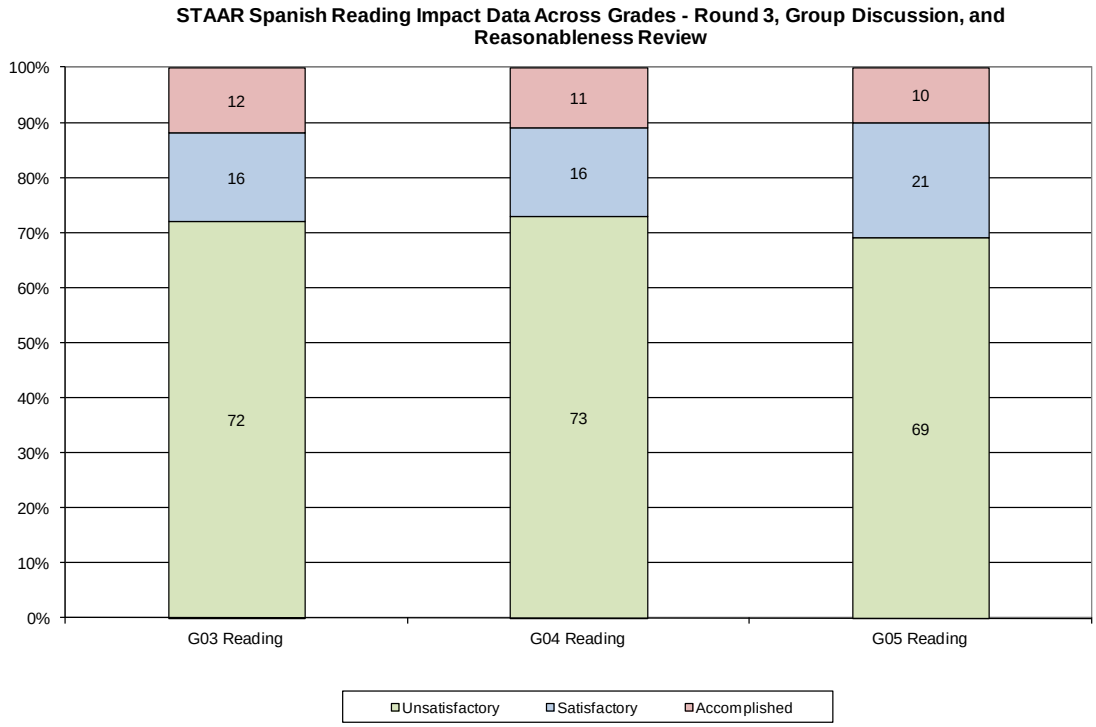
STAAR 3-8 English Reading Impact Data Across Grades - Group Discussion



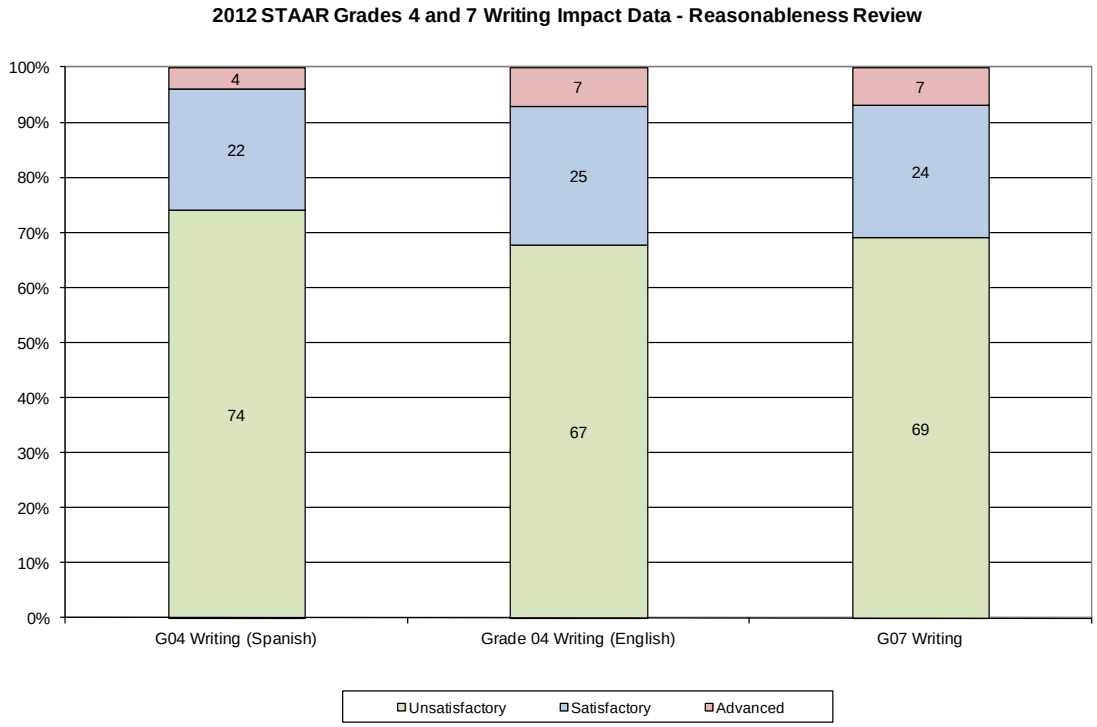
STAAR 3-8 English Reading Impact Data Across Grades - Reasonableness Review



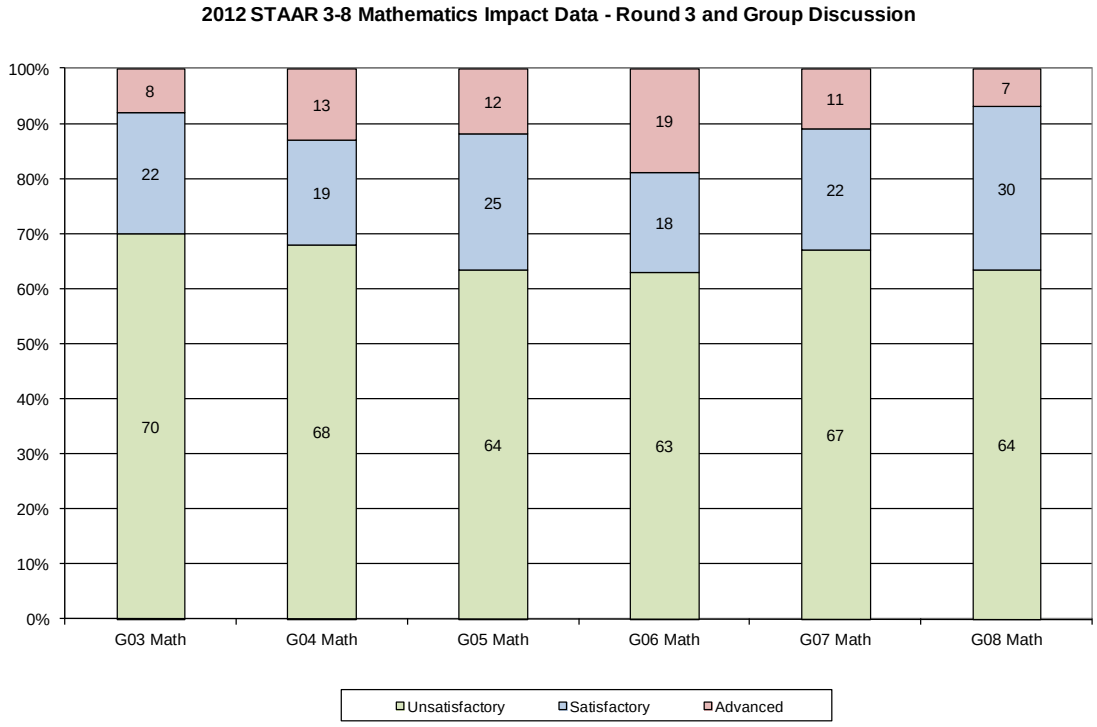
SPANISH READING



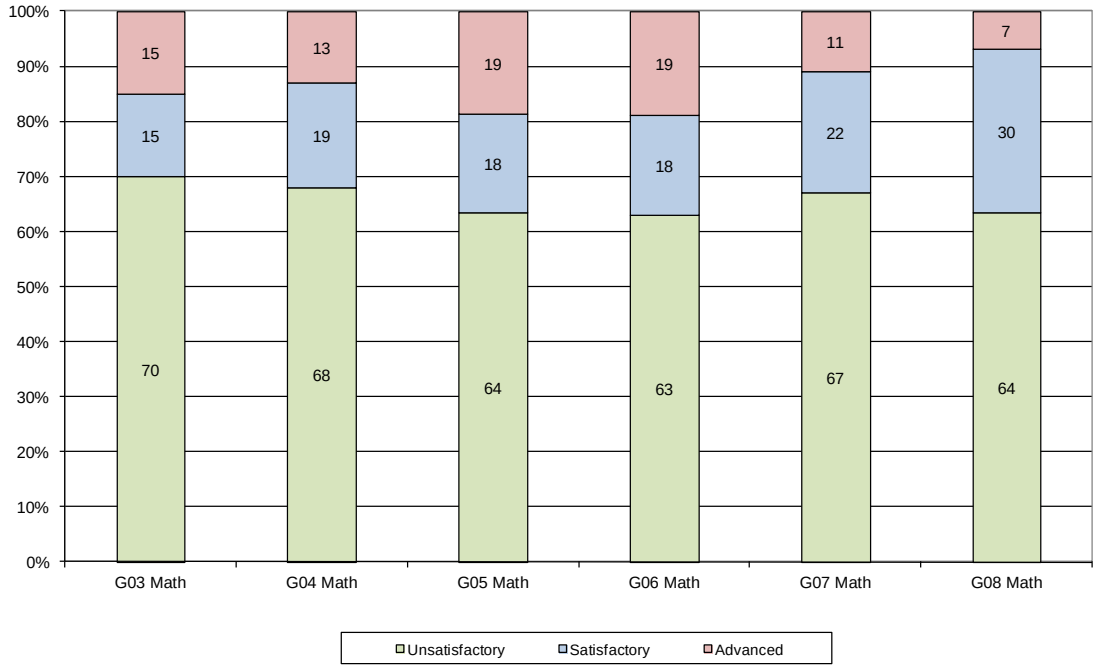
WRITING



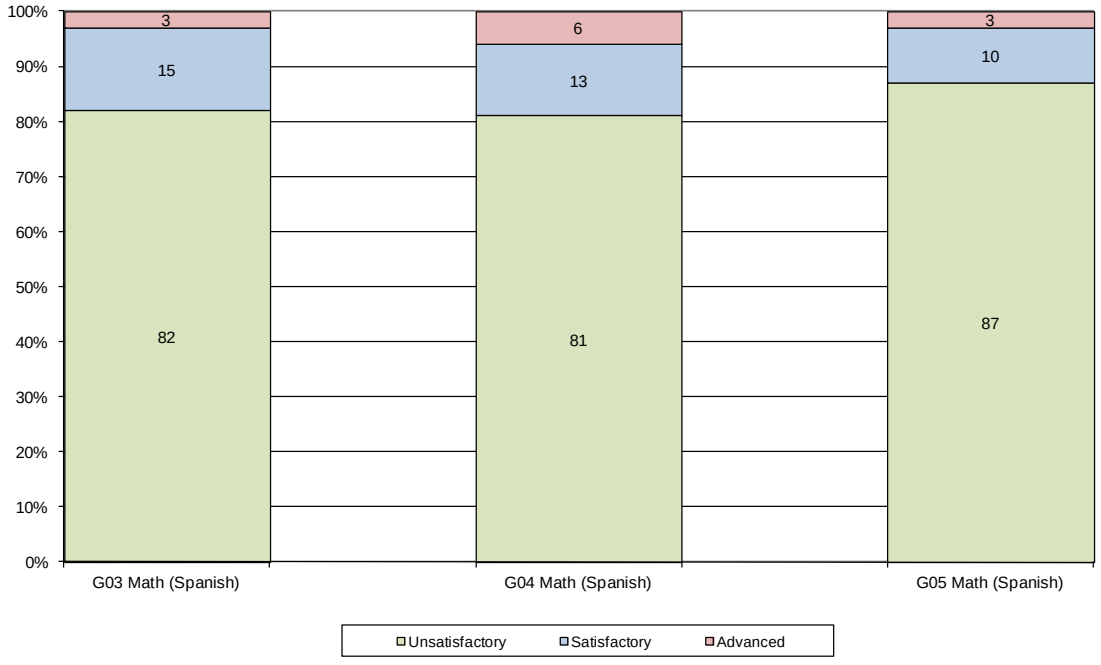
MATHEMATICS

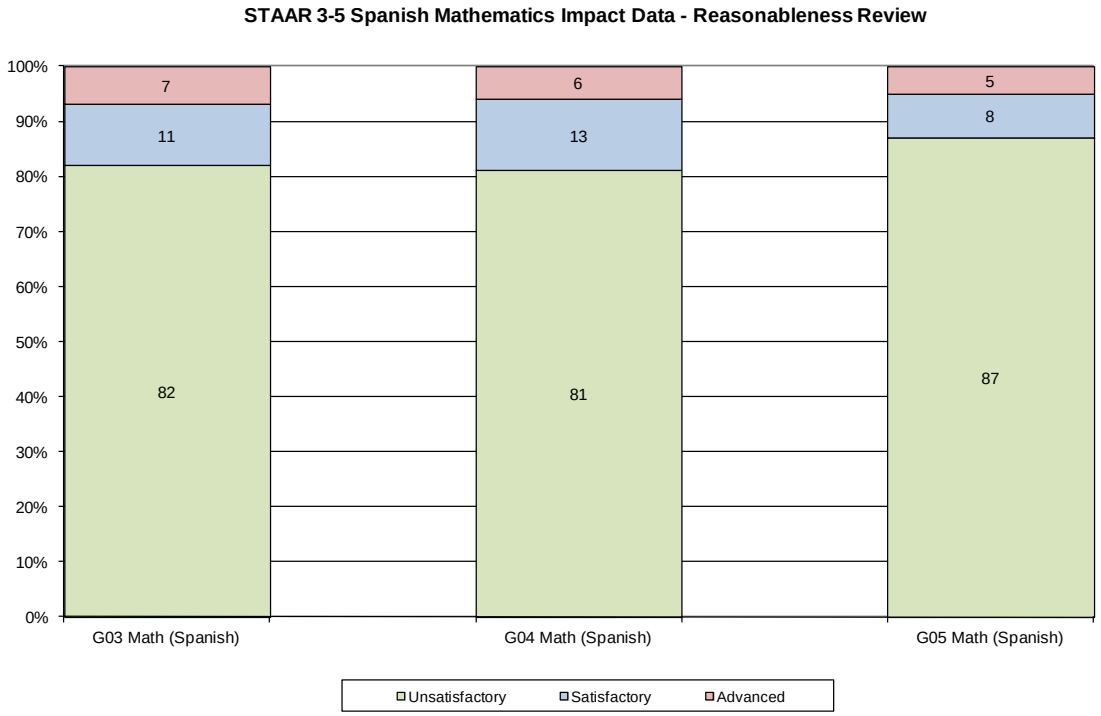


2012 STAAR 3-8 Mathematics Impact Data - Reasonableness Review

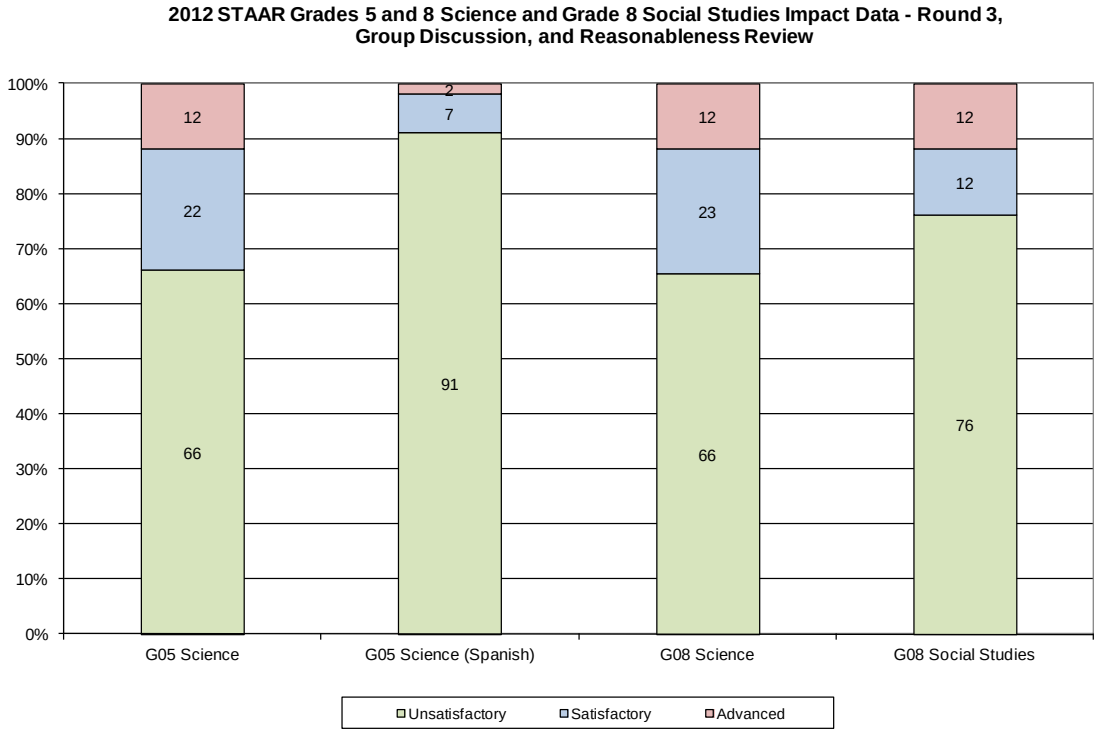


STAAR 3-5 Spanish Mathematics Impact Data - Round 3 and Group Discussion





SCIENCE AND SOCIAL STUDIES



References

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

Bejar, I.I., Braun, H.I., & Tannebaum, R. (2006). *A prospective approach to standard setting.* Paper presenting in *Assessing and modeling development in school: Intellectual growth and standard setting.* College Park, Maryland: University of Maryland.

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.

Embretson, S. E., & Reise, S. R. (2000). *Item Response Theory for psychologists*. New Jersey: Lawrence Erlbaum Associates.

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

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

Haertel, E. H. (2012). The Briefing Book method. In G. J. Cizek (Ed.), *Setting performance standards* (2nd ed.). New York: Routledge.

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.

Impara, J. D., & Plake, B. S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement,* 34*,* 353-366.

Kolen, M.J., & Brennan, R.L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer-Verlag.

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

Linacre, J. M. (2001). *WINSTEPS Rasch Measurement Program, Version 3.32*. Chicago: John M. Linacre.

Livingston, S. & Zieky, M. (1982). Passing scores: a manual for setting standards of performance on educational and occupational tests. Retrieved from ETS Policy and Research Reports website: http://www.ets.org/research/policy_research_reports/passing

Loomis, S.C., & Bourque, M.L. (2001). From tradition to innovation: Standard setting on the National Assessment of Education Progress. In G. J. Cizek (Ed.) *Setting Performance*

Standards: concepts, methods and perspectives. Mahway, NJ: Lawrence Erlbaum Associates.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Mattern, K. D., Patterson, B. F, & Kobrin, J. L. (2012). The validity of SAT scores in predicting first-year mathematics and English grades (College Board Research Report No. 2012-1). New York: The College Board.

Mills, C.N. & Jaeger, R.M. (1998). Creating descriptions of desired student achievement when setting performance standards. In L. Hansche (Ed.), *Handbook for the development of performance standards: Meeting the requirements of Title I.* Washington DC: CCSSO.

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. (pp.301–322). New York: Routledge.

Perie, M. (2008). *A Guide to Understanding and Developing Performance-Level Descriptors.* Hoboken, NJ: National Council on Measurement in Education, Wiley.

Perie, M., Hess, K., & Gong, B. (2008). *Writing Performance Level Descriptors: Applying lessons Learned from the General Assessment to Alternate Assessments based on Alternate and Modified Standards.* Paper presented at the Annual Meeting of the National Council of Measurement in Education (NCME). New York, NY.

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

Phillips, S. E. (2002). *Setting Standards on the TAKS Test: a Modified Item Mapping Procedure*. BETA, Inc., NCS Pearson, and TEA.

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.

Rasch, G. (1966). An Individualistic Approach to Item Analysis. In *Readings in Mathematical Social Science*, edited by Paul F. Lazarfeld and Neil W. Henry. Chicago, IL: Science Research Associates.

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

Redfield, D., & Sheinker, J. (2006). *Comprehensive Academic Systems Alignment and Calibration.* Presentation for Edvantia.org.

Wright, B. D. & Masters, G. N. (1982). *Rating Scale Analysis*. Chicago, IL: Mesa Press.

Wright, B. D. & Stone, M. H. (1979). *Best test design: Rasch measurement*. Chicago, IL: Mesa Press.

Zieky, M. J., & Livingston, S. A. (1977). *Manual for setting standards on the basic skills assessment tests*. Princeton, NJ: Educational Testing Service.