New York State Regents Examination in Algebra 2/Trigonometry

Standard Setting Technical Report

Prepared for the
New York State Education Department by Pearson
July 29, 2010

Table of Contents

List of Tables

List of Figures

Executive Summary

The standard setting process for the New York State Regents Examination in Algebra 2/Trigonometry consisted of three activities: the Pre-Policy Measurement Review Panel meeting, the Item Mapping Standard Setting meeting, and the Post-Policy Measurement Review Panel meeting. This document provides a detailed description of each of these activities. The main purpose of these standard setting activities was to obtain cut score recommendations for the New York State Regents Examination in Algebra 2/Trigonometry. Students could be classified into the following three achievement levels on the assessment: the lowest level, 0–64 (Level 1); 65–84 (Level 2); and the highest level, 85–100 (Level 3).

On April 15, 2010, a Pre-Policy Measurement Review meeting was conducted in Albany, New York. This meeting was convened to provide recommendations for the acceptable percentage of New York State students who should be classified in each achievement level on the New York State Regents Examination in Algebra 2/Trigonometry.

On June 21-22, 2010, an item mapping standard setting meeting was conducted using two committees. In the afternoon of June 22, 2010, selected members of the two committees also formed a synthesis group to reconcile the recommendations from the two independent committees. The purpose of this meeting was to recommend cut scores based on the content standards and achievement level descriptors for the same assessment.

Finally, in the evening of June 22, a Post-Policy Measurement Review Panel meeting was conducted. This meeting, which included panelists from the Pre-Policy Measurement Review Panel, integrated results from the Pre-Policy Measurement Review Panel meeting and the Item Mapping Standard Setting meeting.

In this technical report, panelists, materials, methodologies, and results are presented for each of the three stages for the standard setting activity for the New York State Regents Examination in Algebra 2/Trigonometry. A separate executive summary was provided to the state the day following the standard setting activity outlining the methodologies and major findings. More details are provided in the current technical report.

Pre-Policy Measurement Review Panel

On Thursday, April 15, 2010, the New York State Education Department (NYSED) conducted the Pre-Policy Measurement Review meeting in Albany, New York. This meeting was convened to provide recommendations for the expected percentage of New York State students who should be classified in each achievement level on the New York State Regents Examination in Algebra 2/Trigonometry. During this one-day meeting, panelists participated in two rounds of discussion in which they were asked to make individual "high" and "low" recommendations as to the expected percentage of students who should be classified in each achievement level. For example, a panel member could recommend that it would be expected if between 25–30 percent of students were classified as *Level 3* (85–100) on the Regents Examination in Algebra 2/Trigonometry. The outcomes of the conference are described in this summary and more detailed information will be provided in a subsequent standard setting technical report.

Panelists

A total of 34 panelists attended. These panelists are policy holders and administrators who were geographically representative of New York State. The panelists represented various stake holder groups such as School Administrators Association of New York State (SAANYS), New York State United Teachers (NYSUT), New York State Council of School Superintendents (NYSCOSS), New York State School Board Association (NYSSBA), Big Five Cities, Special Education Directors, District Superintendents, Assistant Superintendents, Superintendents of Schools, etc. All panelists provided voluntary demographic information. Demographic information from the panelists will be summarized in the subsequent standard setting technical report.

Method and Procedure

The Pre-Policy Measurement Review Panel meeting began with introductions of NYSED staff and the facilitators (Drs. Paul Nichols, Kimberly O'Malley, and Ye Tong). Panelists were then introduced to the purpose of the meeting and the role that they played in the process. Next, Pearson facilitators described the procedure that would be used for the meeting. Panelists then reviewed supporting data, the test history, test design, and impact data for the following assessments:

- New York State Grade 8 Mathematics Test
- New York State Regents Examination in Mathematics A
- New York State Regents Examination in Mathematics B
- New York State Regents Examination in Integrated Algebra
- New York State Regents Examination in Geometry
- National Assessment of Educational Progress (NAEP)
    - National Level Data, Grade 4 Mathematics Test
    - New York State Level Data, Grade 4 Mathematics Test

- National Level Data, Grade 8 Mathematics Test
- New York State Level Data, Grade 8 Mathematics Test

Panelists then broke into three groups. The three groups each reflected the same diversity in geographic representation, experience/expertise, race/ethnicity, etc., as the entire panel. Groups met in separate rooms where they discussed guiding questions and impact data and completed the first round of recommendations before breaking for lunch. Following lunch, the committee reconvened in a single room to review the results from the first round and share information across groups. Then, panelists again broke into three groups and after discussions, they completed a second round of recommendations. The final average recommendations were presented without further discussion in a single room to the whole panel.

The following guiding questions were used at the meeting:

1. What type of differences in impact data do the participants expect across achievement levels?
    - Equal across achievement levels?
    - Increasing across achievement levels?
    - Decreasing across achievement levels?

2. What percentage of students in each achievement level would the panel find acceptable on the new examination?
    - What should be the percentage of students in each achievement level?
    - What variations from these values are acceptable?

3. What, if any, consistency is expected between the data from the current and new testing programs?
    - Should the percentage of students in each achievement level be similar, even if the standards have changed?
    - What differences in impact data between the current and new testing programs are acceptable?

4. What type of consistency in impact data does the panel expect among the Grade 8 Mathematics Test, Regents Examination in Integrated Algebra, Regents Examination in Geometry, and Regents Examination in Algebra 2/Trigonometry?
    - What are the differences in impact data among the testing programs?
    - Should the percentage of students in each achievement level be similar even though the tests measure different knowledge and skills?
    - What differences among the programs are acceptable?

5. What, if any, consistency is expected between national data and New York State?
    - What are the differences between New York State's testing program and NAEP?
    - Should the percentage of students in each achievement level be similar even though the testing programs are not similar?
    - What differences between the results for New York State's testing program and the NAEP testing program are acceptable?

Results

The mean recommended percentage of students was computed by averaging both "low" and "high" recommendations across all panelists. The median was computed in a similar fashion. For Round 2, Table 1 summarizes the panelists' recommendations for the expected percentages of students who should be classified as Level 2 (65–84) and Above and the recommended expected percentages of students who should be classified as Level 3 (85–100). The final recommended expected percentages of students are based on the means of the overall panelists' recommendations.

Table 1. The Round 1 Results for the Expected Recommended Percentage of Students Who Should Be Classified as Level 2 (65–84) and Above or Level 3 (85–100).

		Level 2 & Above	Level 3
Group 1	Mean	73.9	23.4
	Median	75.0	25.0
	Standard Deviation	9.1	4.9
	Minimum	45.0	15.0
	Maximum	90.0	32.0
Group 2	Mean	74.5	22.4
	Median	77.5	22.5
	Standard Deviation	15.7	9.1
	Minimum	25.0	5.0
	Maximum	95.0	40.0
Group 3	Mean	75.6	25.6
	Median	75.0	25.0
	Standard Deviation	7.4	8.8
	Minimum	65.0	5.0
	Maximum	90.0	50.0
Committee	Mean	74.7	23.9
	Median	75.0	25.0
	Standard Deviation	10.9	7.8
	Minimum	25.0	5.0
	Maximum	95.0	50.0

Table 2. The Round 2 Results for the Expected Recommended Percentage of Students Who Should Be Classified as Level 2 (65–84) and Above or Level 3 (85–100).

		Level 2 & Above	Level 3
Group 1	Mean	76.5	23.3
	Median	80.0	22.5
	Standard Deviation	8.0	5.0
	Minimum	55.0	15.0
	Maximum	90.0	35.0
Group 2	Mean	81.2	23.0
	Median	84.0	25.0
	Standard Deviation	10.2	9.2
	Minimum	60.0	5.0
	Maximum	95.0	35.0
Group 3	Mean	77.9	24.2
	Median	77.5	22.5
	Standard Deviation	8.8	8.8
	Minimum	60.0	10.0
	Maximum	90.0	40.0
Committee	Mean	78.4	23.5
	Median	80.0	25.0
	Standard Deviation	9.1	7.7
	Minimum	55.0	5.0
	Maximum	95.0	40.0

Figures 1 and 2 show the percentages of students in each achievement level using the mean recommendation across all panelists from the two rounds.

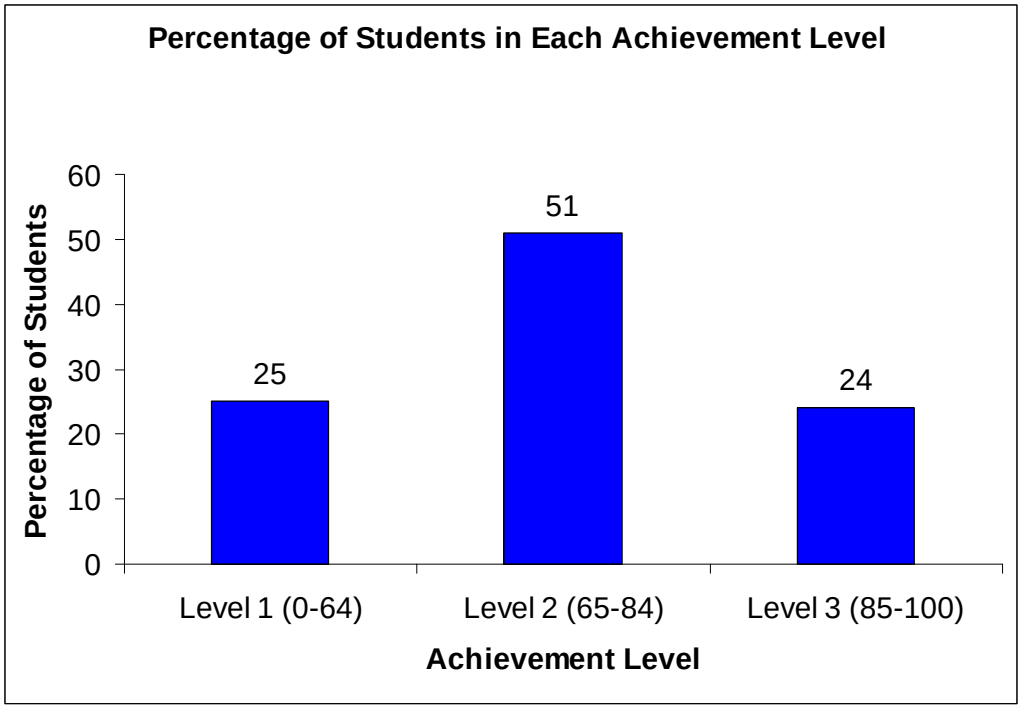

Figure 1. The Mean Percentages of Students in Each Achievement Level Recommended by the Panelists Following <u>Round 1</u>.

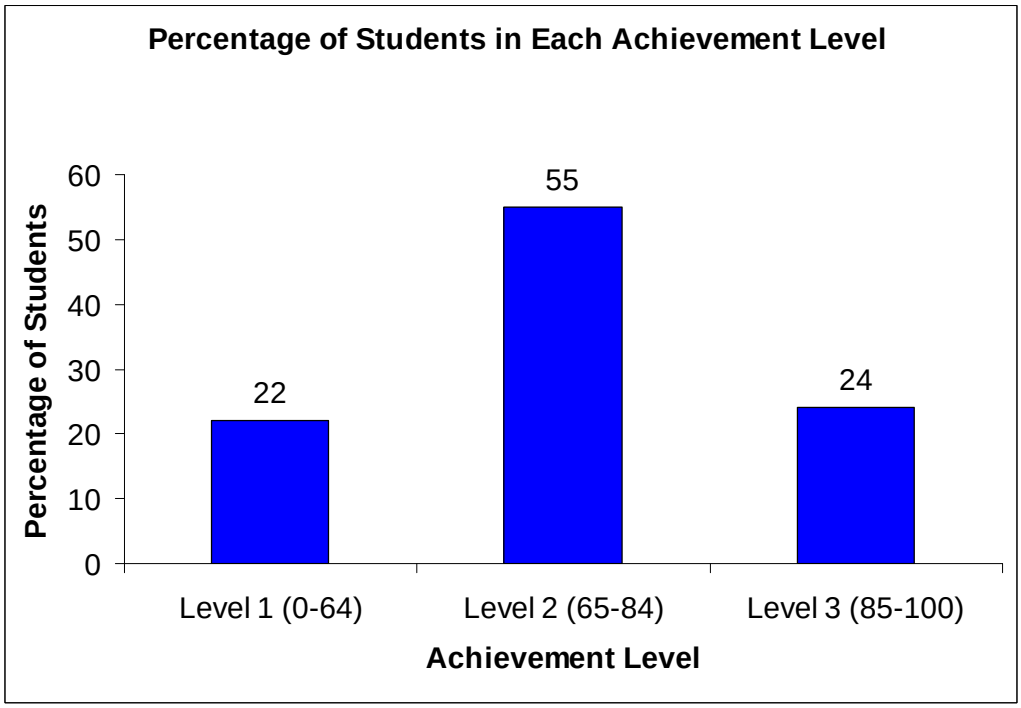

Figure 2. The Mean Percentages of Students in Each Achievement Level Recommended by the Panelists Following <u>Round 2</u>.

Round 2 results were presented to the panelists and were considered the final recommendations from the Pre-Policy Measurement Review meeting. All of the materials used at the meeting were collected and boxed and were made available on June 22, 2010, when the Post-Policy Measurement Review meeting was held after the Item Mapping Standard Setting of the New York State Regents Examination in Algebra 2/Trigonometry.

Evaluations

Exit surveys were administered following the completion of the Pre-Policy Measurement Review Panel meeting. At the end of the meeting, all participants completed the exit survey. Panelists answered each question by choosing one of the following: "totally disagree," "disagree," "neutral," "agree" and "totally agree". For ease of summary, in Table 3, a scale from 1 to 5 was used, with 1 representing "totally disagree" and 5 representing **"**totally agree." The survey questions and the results are presented in Table 3. As can be observed from the summary of the exit survey results, the participants generally had very positive feedback about the process and the outcome of the meeting.

Table 3. The Questionnaire Results for the Pre-Policy Measurement Review Panel Meeting.

Question	Mean	Median	Maximum	Minimum
1. The method for making recommendations on the expected percent of students who should be classified in each achievement level was conceptually clear.	4.15	4	5	3
2. I had a good understanding of the design of the New York State Regents Examination in Algebra 2/Trigonometry.	3.94	4	5	2
3. I had a good understanding of the design for the other assessments presented.	4.15	4	5	2
4. After the first round of ratings, I felt comfortable with the method for making recommendations.	4.32	4	5	2
5. After the second round of ratings, I felt comfortable with the method for making recommendations.	4.44	4	5	3
6. I found the feedback on the recommendations of other panelists useful in making my own recommendations.	4.41	4	5	4
7. I found the feedback on the overall group recommendations useful in making my own recommendations.	4.26	4	5	3
8. I feel confident that the final cut score recommendations reflect the achievement levels associated with the Regents Examination in Algebra 2/Trigonometry.	4.25	4	5	3

Item Mapping Standard Setting

Two committees of New York State educators convened June 21–22, 2010, in Albany, New York, to recommend standards for the New York State Regents Examination in Algebra 2/Trigonometry. The first committee, Committee A, had 29 educators and the second committee, Committee B, had 28 educators. The item mapping procedure was applied to recommend the cut scores.

Panelists

All panelists provided voluntary demographic information. Table 4 presents a summary of gender representation across both committees; Table 5 provides a summary of the ethnic representation of both committees; Table 6 lists the distribution of geographic locations of the panelists; and Table 7 summarizes the educational experience distribution between the two committees.

Table 4. Number of Male and Female Panelists in Committees A and B

	Committee A	Committee B
Female	21	16
Male	8	12

Table 5. Summary of the Ethnic Representation of the Panelists in Committees A and B

	Committee A	Committee B
White	18	12
Hispanic	1	4
African American	4	5
Asian	4	3
Missing	2	4

Table 6. Distribution of Geographic Locations of Panelists for Standard Setting

	Committee A	Committee B
North Country	2	3
Long Island	2	2
NYC	7	7
Lower and Mid Hudson Valley	3	4
Capital Region	4	3
Central NY	3	2
Western NY	8	7

Table 7. Education Roles of Panelists in Committees A and B

	Committee A	Committee B
Mathematics Teachers	26	23
Special Education Teachers	15	11
Bilingual Teachers	12	8
Curriculum/Department/Test Coordinator	1	1
Math Department Chair	2	4

Method

Panelists used an item mapping methodology, sometimes referred to as a bookmark approach, to recommend standards of the Regents Examination in Algebra 2/Trigonometry. The item mapping methodology is typically conducted by using the following materials:

- Achievement level descriptors (ALDs)
- Ordered item books
- Item map

A description of each of these is provided to give background for a description of the item mapping methodology. Following the description of these materials, a description of the typical item mapping methodology is presented.

Achievement Level Descriptors

Standard setting panelists are tasked with estimating the performance of a group of students; e.g., the Basic, Proficient, or Advanced student. Students are grouped into these achievement levels as a way to establish and communicate achievement goals. The achievement levels define what students should know and be able to do when they have reached these achievement levels. For example, what should a student who has reached the Proficient level know and be able to do? States or other test developers create descriptions of what students should know and be able to do at different achievement levels. These descriptions are called achievement level descriptors (ALDs).

Generally, achievement levels represent a broad range of achievement. For example, more than one fourth of the students in a grade level for a state may be classified as failing within the Basic achievement level.

The general ALDs that attempt to capture the range of achievement represented by achievement levels are too vague for standard setting panelists tasked with estimating the performance of students in each achievement level. Panelists make ratings of items, student work samples, or students, using descriptions of what students know and can do at each achievement level. Panelists need descriptions that contain enough detail to support reliable ratings both within panelists, across occasions, and across panelists.

To support reliable ratings in standard setting, descriptions of what just Proficient or just Advanced students know and can do are created. These students that are just Proficient or just Advanced are known as threshold examinees because they define the threshold of the achievement level. Threshold examinees are students with the minimum level of proficiency needed to make it into a particular achievement level.

The descriptions of what just Proficient or just Advanced students know and can do play a central role in standard setting. The panelists are instructed to use these ALDs of what just Proficient or just Advanced students know and can do as the frame of reference for each judgment. The construct being measured is the panelists' representation of just Proficient or just Advanced students' performance. The measurement of that construct results in cut points recommended by panelists.

The logic of using ALDs for threshold students to delimit the range of achievement represented by achievement levels is straightforward. The ALDs for threshold students describe what the most minimally qualified student in that achievement level knows and can do. Students who are not likely to know or be able to do what the threshold students know and can do must fall into the previous achievement level. Students who are likely to know or be able to do more than what the threshold students know and can do must fall into the current or succeeding achievement levels.

Ordered Item Book

Under the item mapping method, panelists review test items from least to most difficult. Panelists are typically given a book of test items, called an ordered item book, to help them with this review. The items in this book are presented one item per page and are ordered from the least difficult items to the most difficult. Often, a three-ring binder is used for the ordered item book.

The ordered item book may include both selected-response and constructed-response items. Each selected-response item, such as a true/false item or a multiple-choice item, is presented only once in the book. A multiple-choice item page will show the test item stem and alternatives, as well as the correct response. A true/false item page will show the test item and the correct response.

Each constructed-response item is presented multiple times, corresponding to the number of score points in the rubric. Each score point for a constructed-response item is presented once in the book, except the 0 score point. For example, a constructed-response item that is scored using a 4-point rubric (0–4) would have four pages in the ordered item book representing score points 1, 2, 3, and 4. The rubric used to score student performance should also be available.

For example, an ordered item book might be constructed for an assessment with 30 multiple-choice items and 8 constructed-response items, each scored on a scale of 1–3. The ordered item book would include 30 pages, 1 page for each of the 30 multiple-choice items. In addition, the ordered item book would include 24 pages, 1 page for each of the three score points for each of the 8 constructed-response items. The ordered item book would total 54 pages.

Sometimes an ordered item book is constructed by using more items than the number of items on an assessment. The items in an ordered item book should represent the categories of content, mix of item formats, and range of difficulty described in the test blueprint. Items from the item bank may be added to provide a better representation of the test blueprint. For example, items from a content category might be added if that category was not fully represented on a test form. Alternatively, items from the item bank may be added so that items represent the entire scale range. For example, the ordered item book may have a sequence of items with difficulty values of 0.00, 0.50, and 1.00 logits. Items with difficulty values near 0.25 and 0.75 logits may be added to the ordered item book to represent the gaps in the scale between items on the test form.

The empirical order of item difficulty must be calculated before the ordered item book can be constructed. Empirical difficulty represents a point on a known ability scale. The ability scale is commonly established by using Item Response Theory under a Rasch or combined model.

Empirical difficulty is calculated for both selected-response and constructed-response items. Selected-response items include true/false items and multiple-choice items. The empirical difficulty for selected-response items is calculated as

the point on the ability scale at which the examinee would have a given probability, called a response probability (RP), of selecting the correct response. Guessing should be factored out of the response probability when computing the empirical difficulty.

Empirical difficulties are computed for those constructed-response items that are scored using a rubric. Constructed-response items are represented by multiple score points, corresponding to the number of score points in the rubric. The empirical difficulty for each score point is calculated as the point on the ability scale at which the examinee would have a given RP of achieving at least that score point. This definition of empirical difficulty for constructed-response score points is conceptually similar to the definition of empirical difficulty for selected-response items. Note that the empirical difficulty should be greater for higher score points than for lower score points. A score point of at least 3 will be more difficult to obtain than a score point of at least 2.

The Regents Examination in Algebra 2/Trigonometry contains 28 selected-response items (multiple-choice items) and 10 constructed-response items. The selected-response items are weighted by 2 for scoring and the constructed-response items are weighted by 1. For the ten constructed-response items, six items have a score range from 0 to 2, three items have a score range from 0 to 3, and one item has a score range from 0 to 6. The raw scores for the Regents Examination in Algebra 2/Trigonometry range from 0 to 86.

Rasch and Partial Credit Models

The Rasch model and the Partial Credit model are used for all the Regents examinations. The Rasch model is applied to fit the multiple-choice items, and the Partial Credit model is applied to fit constructed-response items. Research in standard setting methodology tends to indicate that when an RP value of 0.67 is used, we can achieve the maximum information needed for standard setting. In addition, the RP value of 0.67 has been used historically for other assessments in New York State, such as the Grades 3–8 assessments, this value was also applied when empirical item difficulty of the items was calculated to construct the ordered item book. The Rasch model and the computation of empirical difficulty value with an RP value of 0.67 are discussed below.

When it is a dichotomous item, the Rasch model can be defined as the following.

$$P = \frac{1}{1 + e^{-(\theta - b)}}$$

Using the operational data, item difficulty parameter b was calibrated using WINSTEPS. Based on the theory of the Rasch model, the item difficulty parameter b from the calibration corresponds to a proficiency θ value when the RP value is 0.50. To obtain the item parameter value and hence the corresponding θ value that will have an RP value of 0.67, modification needed to

be conducted on the item parameters. Basically, the following equations needed to be solved for b', the item difficulty, hence the ability level for an RP value of 0.67.

$$0.50 = \frac{1}{1+e^{-(\theta-b)}}$$

$$0.67 = \frac{1}{1+e^{-(\theta-b')}}$$

Solving this equation, we obtained $b' = b + \ln 2 = b + 0.69315$. Therefore, a factor of 0.69315 was added to the multiple-choice item parameters (dichotomous items only) for the items to be included in the ordered item book.

When it is a polytomously scored item (constructed-response item), the formulas are a bit more complicated. The IRT Partial Credit model was used to analyze polytomously scored constructed-response items for the New York State Regents Examinations. The model is defined as

$$P_{xi} = \frac{\exp \sum_{j=0}^{x} (\theta - D_{ij})}{\sum_{k=0}^{m_i} \left[\exp \sum_{j=0}^{k} (\theta - D_{ij}) \right]},$$

where $x = 0,1,...,m_i$. D_{ij} values were available from the calibration of operational data, and they were obtained using a response probability of 0.50, by model definition.

To obtain RP 0.67 difficulty values, more intensive computation needed to be conducted to produce the value. It was more complicated than a simple addition factor, as is the case with dichotomously scored items. The idea was to produce the ability value that would yield a probability of 0.67 for a given score category and above. Basically, the ability value associated with a score value of 2 for a 4-point item indicates the ability that will yield a probability of 0.67 for a student to get a score of at least 2 (including 2, 3, and 4) for this 4-point item. To conduct this computation, an iterative process was employed, with θ in the increment of 0.001, to locate the corresponding b' value that would yield the RP value of 0.67.

b' value was computed for all score points for each of the constructed-response items. Two independent psychometricians conducted the analysis and their results were a 100% match.

After all the values were computed, the ordered book was created by ordering the items in terms of the computed b' values. In addition, items from two anchor forms were included in the ordered item book to include more content and statistical coverage for the test. The ordered item book can be located in Appendix I. There were altogether 99 pages in the ordered item book. The table below indicates the configuration of the ordered item book.

Table 8. Composition of the Ordered Item Book

	Number of Items	Maximum Credit	Number of Pages
Operational Test			
Multiple-Choice	27	1	27
2-credit Item	8	2	16
4-credit Item	3	4	12
6-credit Item	1	6	6
Anchor Forms			
Multiple-Choice	14	1	14
2-credit Item	5	2	10
4-credit Item	2	4	8
6-credit Item	1	6	6

Item Map

The item map is a handout that accompanies the ordered item book and provides additional information for each item. The item map is a table that consists of one row for each item in the ordered item book. The items are listed on the item map in the same order they are presented in the ordered item book; i.e., from least to most difficult based on the empirical item difficulty calculated using an RP value of 0.67. Each row lists information about the item. The following information is commonly provided for each item:

- The page number in the ordered item book
- The original item number on the test form (unless the item is from the test bank)
- The content classification of the item
- The key (unless the row corresponds to a score point for a constructed-response item)
- Maximum score point if the item is a constructed-response item

Following round one of the standard setting procedure, an augmented item map is often distributed to panelists as part of the structured feedback provided between rounds of ratings. The augmented item map presents the information from the original item map and adds information about item difficulty, typically the percentage of students who answered the item correctly (for multiple-choice items) or the percentage of students who earned this score point or higher (for constructed-response items).

Item Mapping Methodology

Under the item mapping standard setting method, panelists are asked to review items in the ordered item book and make a judgment as to the likelihood of threshold examinees answering an item correctly or achieving a given score

point or higher. This judgment is made within a given frame of reference, for a given RP value, and within a given procedure.

The panelists are instructed to use the ALDs as the frame of reference for each judgment. The panelists have completed a warm-up task to become familiar with the ALDs. Sometimes, the panelists may have created the ALDs during an earlier session. These ALDs describe what the threshold examinees at each achievement level (e.g., just Level 2 or Level 3) know and can do. Panelists use only one ALD at a time.

Panelists are instructed to judge the likelihood of threshold examinees answering an item correctly or achieving at a given score point. The RP value used for this assessment was 0.67. Panelists may be instructed to think of this RP value in several ways. Panelists may be instructed to think about a group of 100 threshold students (e.g., just Proficient students). For an RP value of 0.67, panelists are asked to identify the item that 67 of 100 threshold students will answer correctly. Alternatively, panelists may be instructed to think of a typical threshold student, perhaps a student they are teaching or have taught. Again for an RP value of 0.67, panelists are asked to identify the item that this student would have a 67% chance of answering correctly.

The task set for panelists is to read each item or score point in the ordered item book and evaluate the knowledge, skills, and abilities required to respond correctly to the item or to produce a response at the score point. Panelists then compare their evaluation of the cognitive demands of each item and score point to the assigned ALD; e.g., the description of the just Proficient examinees. Panelists should proceed from the least difficult items to most difficult. Keeping in mind the ALD, panelists are instructed to identify the last item or score point that 67 of 100 threshold students should answer correctly. For the item immediately following, panelists should judge that only 66 or fewer of 100 just Proficient examinees would respond correctly. For the item immediately preceding, panelists should judge 68 or more of 100 just Proficient examinees would respond correctly. Panelists then mark the last yes page in the ordered item book, often using a self-adhesive note, and record the item identifier on a record sheet.

Cut Score Computation

The cut score at each achievement level was determined by computing the median from the judge ratings. For a given achievement level, each judge, for each round, had a page number recommendation. These page numbers then were translated into Rasch values where an ability of this level produces an RP value of 0.67 of answering the item correctly. The median of these Rasch values was then computed, which was the cut score recommendation on the θ scale. The raw to θ conversion table was used to look up the corresponding raw score cut. The standard setting θ was likely to be between two θ values on the raw to θ conversion table. To give students the benefit of the doubt, based

on NYSED's direction, the lower of the two θ were identified and its associated raw score was used for the raw score cut recommendation.

This identified raw score represents the minimum raw score that an examinee must attain to be classified into a particular achievement level based on the standard setting methodology. As mentioned before, the ordered item book contained 90 pages representing 90 score points—including both operational items and items from the two anchor forms. The raw to θ conversion table was based on the operational Regents Examination in Algebra 2/Trigonometry and had raw scores ranging from 0 to 86. The median panelist rating was computed for each achievement level. Using that median ability value, the corresponding raw score was identified.

For example, at round three, the median page number for Level 2 in group A was 30. The item on page 30 had the θ value of 0.666 (see Appendix I, ordered item book). Next, we go to the raw to θ conversion table. A raw score of 50 corresponds to a θ value of 0.664; a raw score of 51 corresponds to a θ value of 0.707. Per NYSED's direction, the raw score cut recommended then was 50. The rest of the cut scores were identified using the same algorithm.

Methodological Strengths

The item mapping method has several features that make it an appealing standard setting approach. First, the item mapping method can be used with a mixed-format assessment. Panelists consider both selected-response and constructed-response items when placing bookmarks. Consequently, panelists' cut score recommendations reflect the mix of item formats found on a test.

Second, the task that panelists complete within the item mapping method may be relatively less challenging than the panelists' task under other standard setting methods. Proponents of the item mapping method argue that panelists are required to make relatively few judgments compared with the number of judgments required of panelists under other standard setting methods. For example, panelists using the item mapping method to recommend cut scores for three achievement levels would be required to make only two judgments. In contrast, panelists using an Angoff method to recommend cut scores would be required to make one judgment for each item.

In addition, panelists using the item mapping method are required to spend relatively less time reviewing the test items. A panelist who has reviewed the first group of items and placed the first bookmark need not review those items again to place a subsequent bookmark. The panelist would place the first bookmark and then continue paging through the ordered item book to find the appropriate item on which to place the next bookmark.

Before an item mapping procedure can be conducted, substantial work must be done, including collecting student responses and calibrating and scaling items, using Item Response Theory. Student responses may be collected through either a field test or an operational administration. An operational

administration is likely to provide a larger number of responses, collected under more realistic conditions, than a field test.

Procedure

The standard setting conference began on June 21. The agenda for the standard setting conference is shown in Appendix D. The morning was devoted to introductions of the staff, a description of standard setting, and a description of the Regents Examination in Algebra 2/Trigonometry.

Following the midmorning break, all of the educators remained in the same room and began the process of reviewing ALDs. This activity was recommended by the Technical Advisory Group (TAG) - the two independent committees should discuss ALDs together in one large group prior to the standard setting process. The purpose of the activity was to make sure both committees were using the same expectations for students in each of the achievement levels when recommending achievement standards. This process required several hours and resulted in a set of descriptors for each achievement level (Level 1 (0-64), Level 2 (65-84), and Level 3 (85-100)). Appendix E presents the general ALDs provided by NYSED. The educators then broke into eight small groups and discussed specific ALDs. After the small group discussions, all panelists reconvened, and each group presented its ALDs. A typed summary of the ALDs was captured from the discussions and made available to each of the educators for the rest of the standard setting conference. Appendix E also provides the specific ALDs by the educators. The specific ALDs consisted of two parts: descriptors for each of the three achievement levels, and the most distinguishing features for the students who are at the threshold of each achievement level, Level 2 and Level 3.

After the discussions about ALDs and after the educators agreed on the general expectations of what students should know and be able to do in each of the achievement levels, the large group was broken into two separate committees. From here on, the standard setting process was independent between the two committees.

Each committee met in its own meeting room and began the standard setting process. There were 29 and 28 educators per committee respectively and these educators were pre-assigned to four different tables. A leader was assigned for each table. The item mapping procedure was the methodology used. Panelists were instructed to identify the last item in an ordered item book that a threshold student at a given level would have a response probability of at least 0.67 of answering correctly.

The ordered item books were constructed from operational items from the June 2010 administration and anchor items from a field test administration from 2009. Items were sorted from least to most difficult, using the Rasch item difficulty values based on an RP value of 0.67.

The standard setting process consisted of three rounds of judgments. The ratings sheet used by the panelists is shown in Appendix F.

Panelists were provided with feedback between each round. The feedback was intended to inform the panelists' decisions, but not to dictate their ratings. Following round one, panelists met in small groups of seven or eight panelists. They were provided the cut scores (in terms of ordered item book page number) for each panelist on the basis of the round one ratings in addition to the mean, median, minimum, and maximum cut score at each level for that table. In reviewing the cut score report, individual panelists were asked to think about the following:

- How similar are your cut scores to that of the group (i.e., is a given panelist more lenient or stringent than the other panelists)?
- If so, why is this the case? Do panelists have different conceptualizations of these borderline students? Were ALDs being used when making the ratings?

Panelists were informed that there was no intention for them to come to a consensus on cut score judgments, but they should discuss differences to gain an understanding for why differences exist.

In addition, panelists were provided a list of item *p*-values. Finally, panelists were presented with the raw score cut based on their committee's round one rating. The *p*-values were based on a representative sample of approximately 97,060 students who took the operational exam in June 2010.

Within each committee, panelists were given time to discuss the appropriateness of the committee level cut scores, given the proportion of students that would fall into each level.

Following round two, panelists received the cut scores for each panelist on the basis of the round two ratings, in addition to the mean, median, minimum, and maximum cut score at each level for that table. Next, panelists were given the mean, median, minimum, and maximum cut scores for the committee (across tables). The facilitator led the discussion with all four tables combined and noted the differences and similarities across tables, but reminded the panelists that a consensus was not required.

Next, panelists were provided with the overall cut score on the raw score metric, as well as a graphical display of the percentage of students in each achievement level on the basis of the median cut scores from round two. The impact data was based on the same representative sample that the *p*-values were based on. Panelists were also provided with a graphical display of the percentage of students in each achievement level disaggregated for Grade 9 students and Grade 10 and above students.

Within each committee, the panelists were given time to discuss the appropriateness of the committee level cut scores, given the proportion of students that would fall into each level.

After the panelists had a chance to discuss their current cut score recommendations and the related impact, they provided the rating for round three, the final round. The median from round three ratings from each committee was considered the final cut score recommendation for the committee.

After round three rating and analysis, both committees reconvened. The final round recommendations from both committees, along with their impact, were presented. Next, the panelists were instructed to fine tune the ALDs they had developed in the first day, prior to the standard setting activity. The edits and the final ALDs were captured. They are provided in Appendix E.

After completion of the editing of ALDs, the panelists filled out exit surveys, were thanked for their time and participation, and were dismissed. The table leaders from each committee, a total of eight people, were asked to stay and participate in the synthesis meeting. The synthesis meeting was scheduled based on the advice from TAG, and the purpose of the synthesis was to focus on the differences in the cut score recommendations from the two independent committees. In fact, after round three, the two committees provided exactly the same cut score recommendations, as presented in the following results section. Still, the synthesis group met and focused on the items that were around the cut score recommendations. These eight panelists focused on the knowledge and skills those items were measuring, and how they related to the ALDs and especially the differences between the students who were just below the achievement level and the students who were just above the achievement level (the threshold students). The synthesis group then made their final recommendation.

Results

Table 9 summarizes cut score recommendations in terms of page number as well as raw score cuts for achievement Level 2 and Level 3 for round one. Table 10 summarizes cut score recommendations for round two, and Table 11 presents the final cut score recommendations for round three. For each round, the mean, median, minimum, and maximum page number recommendations are presented, as well as the raw score cut recommendation based on the median recommendation from the entire committee. As can be observed from the tables for each round, the cut score recommendations on the raw score metric were very consistent between the two independent committees. Discussions on ALDs with the two committees combined probably contributed to the consistency between the two committees.

Comparisons across rounds also indicate that the cut score recommendations did not fluctuate much between rounds—basically around 1 or 2 points on the raw score metric. Item empirical difficulty (p values) were presented after round one and impact data (percentage of students in each achievement level based on the cut score recommendation) was presented after round two. These two pieces of additional information seemed to have no great effect on the overall cut score recommendations in either of the two committees. Standard deviations are

not presented in these tables because, as the previous section indicated, all the computations were conducted at the θ metric and translated back to either page number or raw scores. With mean, median, minimum, and maximum values, the translation worked well; but with standard deviation, the translation would not have worked well. Therefore, standard deviations were not provided.

Table 9. Cut Score Recommendations by Committee for Level 2 (65–84) and Level 3 (85–100), Round 1

		Page Number				Raw Score
		Mean	Median	Minimum	Maximum	
Committee A	Level 2	33	28	12	94	47
	Level 3	79	75	49	97	65
Committee B	Level 2	34	34	9	69	50
	Level 3	81	81	50	96	67

Table 10. Cut Score Recommendations by Committee for Level 2 (65–84) and Level 3 (85–100), Round 2

		Page Number				Raw Score
		Mean	Median	Minimum	Maximum	
Committee A	Level 2	29	27	15	51	47
	Level 3	73	74	36	86	64
Committee B	Level 2	31	30	22	47	48
	Level 3	76	75	66	90	65

Table 11. Cut Score Recommendations by Committee for Level 2 (65–84) and Level 3 (85–100), Round 3

		Page Number				Raw Score
		Mean	Median	Minimum	Maximum	
Committee A	Level 2	26	25	18	48	45
	Level 3	73	74	53	88	64
Committee B	Level 2	29	30	19	45	48
	Level 3	75	75	58	90	65

Figure 3 and Figure 4 present the percentage of students in each achievement level using the cut score recommendations after rounds two and three. The impact data were based on a representative sample of 97,060 students who participated in the operational testing of the June 2010 Algebra 2/Trigonometry administration. These figures were presented to the panelists after round two and round three, respectively. In order to keep the two committees totally independent during the standard setting process, each committee was only presented the impact data based on their own recommendation.

Not surprisingly, the two committees had exactly the same impact data based on the round three rating because their cut score recommendations were identical.

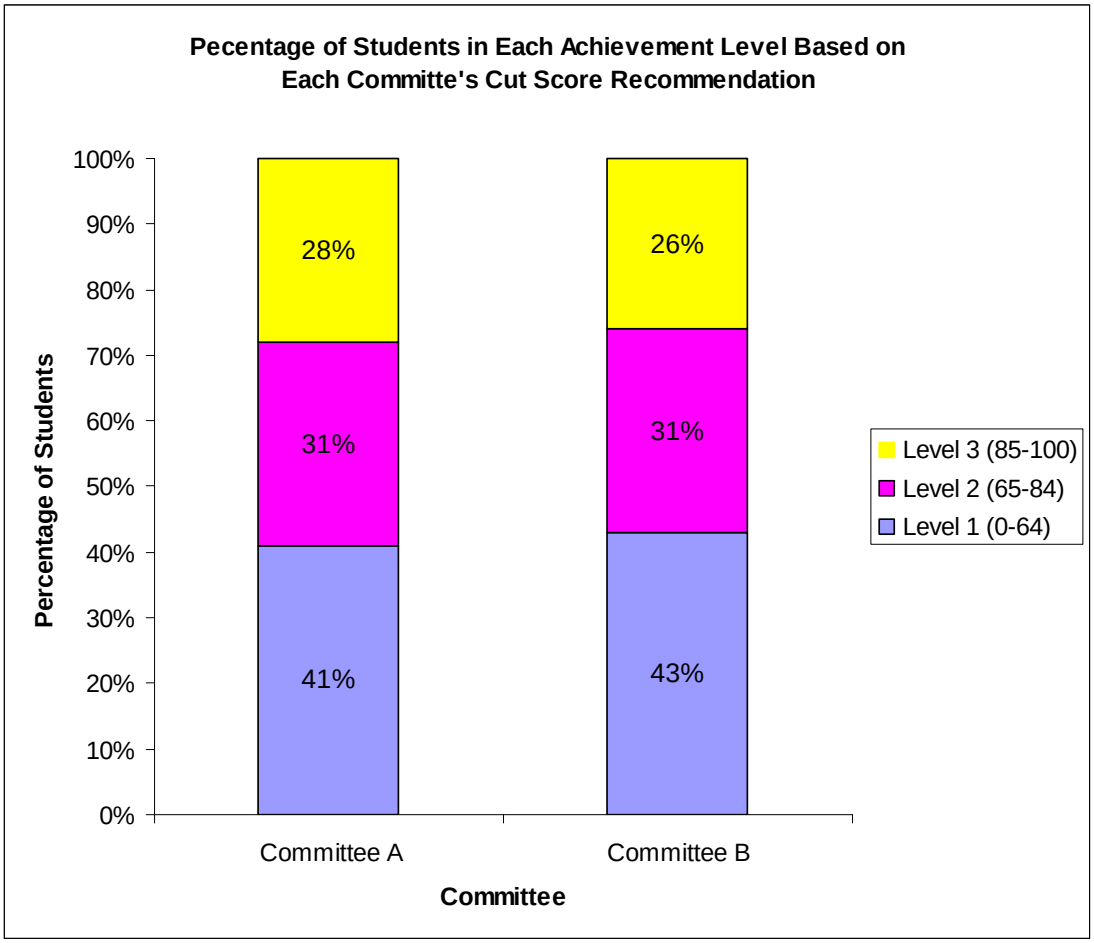

Figure 3. The Percentage of Students in Each Achievement Level Using Cut Score Recommendations by Committee after Round 2

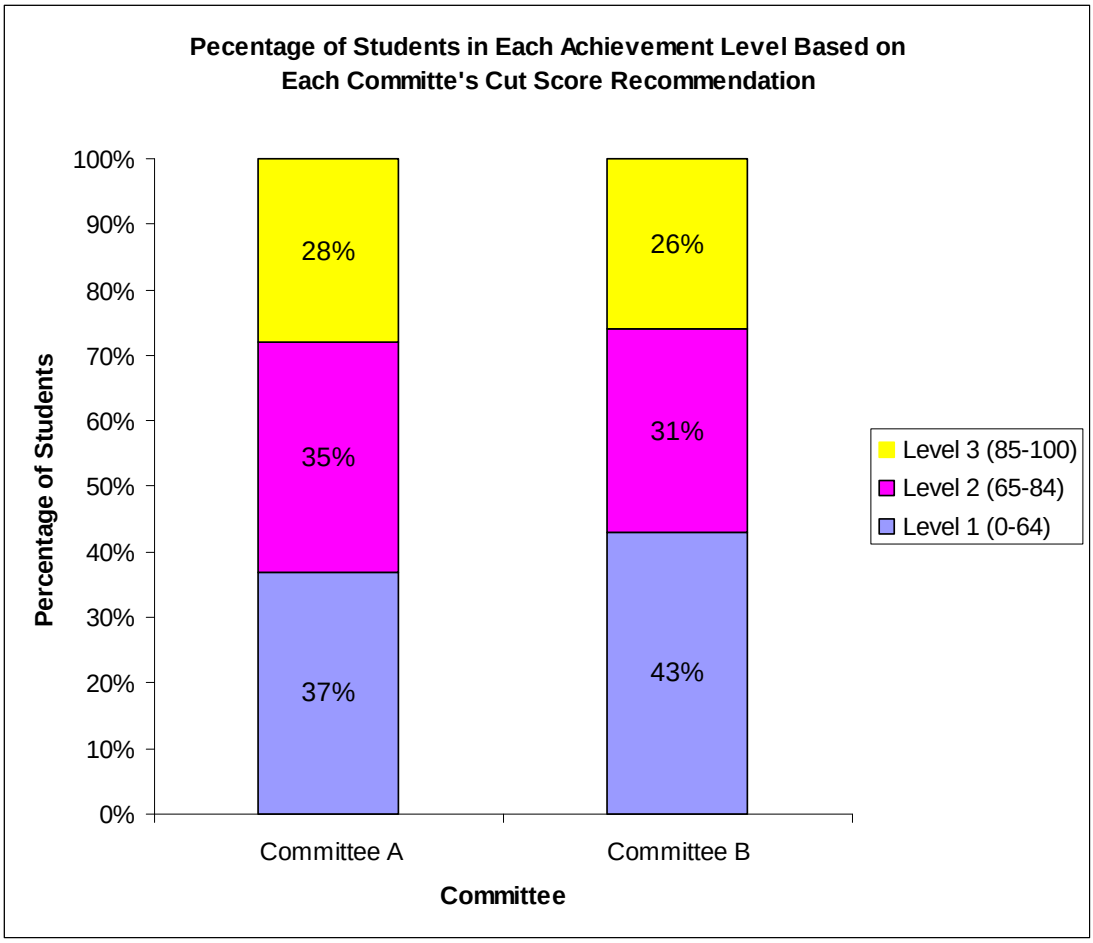

Figure 4. The Percentage of Students in Each Achievement Level Using Cut Score Recommendations by Committee after Round 3

After the two independent standard-setting committees A and B completed their final recommendations and finalized the ALDs, a synthesis group was convened. The table leaders were invited to stay to participate in the synthesis group. There were eight table leaders who participated in this activity. The purpose of the synthesis was to further examine the differences between the cut score recommendations of the two independent committees, if there were any, to discuss the differences, and to come to a final recommendation.

One panelist from each committee gave a brief description of the recommendations and the rationale behind the recommendations. Next, the panelists were asked to observe the cut score recommendations from both committees and to use the ordered item book to further look at the items identified as the bookmarks. The panelists were asked to observe the knowledge and skills the items at the cut and around the cut were measuring, the related ALDs for borderline "just make it" students for each achievement level, and to make an overall recommendation.

Only one round of rating was conducted for the synthesis meeting, with each panelist participating in the synthesis providing a rating for each of the two cuts: Level 2 (65–84) and Level 3 (85–100). Table 12 summarizes the results from the synthesis meeting.

Table 12. Cut Score Recommendation from the Synthesis Meeting

	Page Number				Raw Score
	Mean	Median	Minimum	Maximum	
Level 2	26	26	24	30	46
Level 3	75	75	75	76	65

Panelist Variability

In order to describe the variability in panelists' judgments, a Generalizability Theory (G-Theory) study was performed. This information could be used to determine how similar the cut scores might be if a different set of panelists or different composition of small groups was used to set cut scores. For this investigation, the sources of variability of interest were panelists, small groups, and rounds. For each cut score, the variance associated with each of these sources was estimated using the maximum likelihood SAS VARCOMP procedure. For this study, the number of rounds was treated as a fixed factor (3 rounds in total, a typical practice in standard setting meetings), meaning that if the standard-setting meeting was held again, the same number of rounds would be used. In addition, because judges discussed all activities in small groups, their judgments were considered dependent on group membership. Therefore, judges were considered "nested" within tables.

The judge variability estimates based on the generalizability theory are presented in Table 13 through Table 16.

Table 13. Generalizability Theory Analysis of Judge Variability, Level 2 Cut Score, Committee A

Variance Component	Estimated Variance Component	Applied Variance Component	Percent of Variance	Error Variance	Standard Error
Table	21.1041	21.1041	44		
Judge:Table	1.0440	1.0440	2		
Round	0.1936	0.1936	0		
Table x Round	2.5980	2.5980	5		
Remaining	23.1050	23.1050	48		
				5.35411	2.31389

Table 14. Generalizability Theory Analysis of Judge Variability, Level 3 Cut Score, Committee A

Variance Component	Estimated Variance Component	Applied Variance Component	Percent of Variance	Error Variance	Standard Error
Table	5.9720	5.9720	23		
Judge:Table	0.5981	0.5981	2		
Round	1.1096	1.1096	4		
Table x Round	0.8113	0.8113	3		
Remaining	16.9526	16.9526	67		
				1.54879	1.24450

Table 15. Generalizability Theory Analysis of Judge Variability, Level 2 Cut Score, Committee B

Variance Component	Estimated Variance Component	Applied Variance Component	Percent of Variance	Error Variance	Standard Error
Table	7.4833	7.4833	29		
Judge:Table	3.6540	3.6540	14		
Round	2.7557	2.7557	11		
Table x Round	-0.1098	0.0000	0		
Remaining	11.4850	11.4850	45		
				1.93764	1.39199

Table 16. Generalizability Theory Analysis of Judge Variability, Level 3 Cut Score, Committee B

Variance Component	Estimated Variance Component	Applied Variance Component	Percent of Variance	Error Variance	Standard Error
Table	1.62386	1.62386	11		
Judge:Table	3.40972	3.40972	22		
Round	0.89253	0.89253	6		
Table x Round	-0.40245	0.00000	0		
Remaining	9.29563	9.29563	61		
				0.46408	0.68123

Evaluations

An exit survey was completed by each panelist following the completion of standard setting. Panelists answered each question, using a scale of 1–5, 1 being "**totally disagree**" and 5 being "**totally agree**." The survey questions and the results for are shown in Table 17.

Table 17. Questionnaire Results for Both Committees

Question	Mean	Median	Minimum	Maximum	Standard Deviation
1. The method for providing the rating was conceptually clear.	4.3	4.0	2.0	5.0	0.7
2. I had a good understanding of what the test was intended to measure.	4.3	4.0	3.0	5.0	0.7
3. I could clearly distinguish between student achievement levels.	4.1	4.0	2.0	5.0	0.8
4. After the first round of ratings, I felt comfortable with the standard setting procedure.	4.3	4.0	1.0	5.0	0.9
5. I found the feedback on p-values useful.	4.0	4.0	3.0	5.0	0.7
6. I found the feedback reports on the rating of panelists useful.	4.3	4.0	2.0	5.0	0.7
7. I found the feedback on the percentage of the students tested that would be classified at each achievement level useful.	4.2	4.0	1.0	5.0	1.0
8. Table discussion was open and honest.	4.7	5.0	2.0	5.0	0.6
9. I believe that my opinions were considered and valued by my group.	4.7	5.0	3.0	5.0	0.6
10. I am confident that my round 3 ratings for 65-84 and 85-100 reflect the knowledge, skills, and abilities described in the achievement level descriptors.	4.5	5.0	1.0	5.0	0.7
11. I am confident that the final cut score recommendations reflect the achievement levels associated with the New York State Regents Examination in Algebra 2/Trigonometry.	4.0	4.0	1.0	5.0	1.1
12. I would defend the standards recommended by our committee.	4.3	5.0	1.0	5.0	1.1

A decision factor survey was also completed by each panelist following the completion of standard setting. Panelists answered each question, using a scale of 1–5, 1 being "not at all" and 5 being "very strongly." The decision factor survey questions and the results for Committee A are shown in Table 18. As can be observed from the tables, generally speaking, the most influential factors in panelists' decision making during the recommendations at standard setting appear to be their experience in education and their understanding of the ALDs.

Table 18. Decision Factor Survey Results

Decision Factors	Mean	Median	Min	Max	Standard Deviation
1. Your experience in education	4.4	5.0	2.0	5.0	0.8
2. Prior to this item mapping standard setting, your perceptions about students in each of the three achievement levels	3.4	3.0	1.0	5.0	0.9
3. Your prior knowledge about standard setting	2.5	2.0	1.0	5.0	1.5
4. The orientation on standard setting presented today	3.7	4.0	1.0	5.0	1.0
5. Your perception of the high stakes versus low stakes context of the New York State Regents Examination in Algebra 2/Trigonometry	3.2	3.0	1.0	5.0	1.1
6. Your thinking about students in each achievement level with whom you have had experience	3.7	4.0	1.0	5.0	1.1
7. The consequences of your decisions for NCLB	1.8	1.0	1.0	5.0	1.2
8. Your concerns about district or state political or economic issues	2.3	2.0	1.0	5.0	1.3
9. Your understanding of the achievement level descriptors	4.0	4.0	2.0	5.0	0.7
10. The item p values that were presented after round 1	3.2	3.0	1.0	5.0	1.0
11. The impact data presented after rounds 1 and 2	3.2	3.0	1.0	5.0	1.1
12. The feedback report on estimated raw score cuts from rounds 1 and 2	3.3	3.0	1.0	5.0	1.1
13. Your interactions with your fellow panelists in your group before **round 1**	3.2	3.0	1.0	5.0	1.3
14. Your interactions with your fellow panelists in your group before **round 2**	3.5	3.0	2.0	5.0	1.0
15. Your interactions with your fellow panelists in your group before **round 3**	3.4	3.0	1.0	5.0	1.3
16. Your interactions with your fellow panelists in the large group discussion	3.2	3.0	1.0	5.0	1.2

Post-Policy Measurement Review Panel

The Post-Policy Measurement Review Panel met on the afternoon of Tuesday, June 22, following the completion of the item mapping committee meetings. Both the item mapping meeting and the Post-Policy Measurement Review Panel meeting were held in Albany. The Post-Policy Measurement Review Panel was convened with panelists from the Pre-Policy Measurement Review Panel. The purpose of the Post-Policy Measurement Review Panel was to integrate results from the Pre-Policy Measurement Review Panel meeting and the two committees from the item mapping meeting.

Panelists

This meeting was convened with 30 panelists from the Pre-Policy Measurement Review Panel. Four participants from the Pre-Policy Measurement Review Panel did not attend the meeting.

Method and Procedure

The Post-Policy Measurement Review Panel meeting began with introductions of the facilitator and NYSED staff. Panelists were then introduced to the purpose of the meeting. Panelists were instructed that they were to review and integrate results from the Pre-Policy Measurement Review and the Item Mapping Standard Setting meetings. The product of this activity would be final recommendations for the percentage of students in each achievement level that reflects the influence of both meetings.

Following these initial activities, panelists reviewed results from the Pre-Policy Measurement Review Panel. They were also given an explanation of the item mapping methodology. They then reviewed the results for committees A and B from the Item Mapping Standard Setting and Synthesis meetings.

Following the review of the methods and results of previous meetings, panelists were asked to try to independently integrate results from both meetings. They then discussed the integration of these results. Finally, the panelists made independent recommendations as to the percentage of students in each achievement level.

Following these independent recommendations, the panelists were presented with the mean, median, minimum, and maximum percentage of students in each achievement level for the committee. They were asked to share with the rest of the committee how they integrated the results from the previous meetings.

Results

Table 19 summarizes, for the Post-Policy Measurement Review Panel, the panelists' recommendations for the percentage of students who should be classified as Level 2 and above and the percentage of students who should be classified as Level 3. Table 20 presents the final round of recommendations from the post-policy meeting. The panelists requested that the median value be used for the overall recommendation due to outliers. Raw score cuts corresponding to the median recommendation from both rounds were presented to the panelists during the meeting.

Figure 5 and Figure 6 show the percentage of students in each achievement level using mean recommendations from the Post-Policy Measurement Review Panel for round 1 and the final round. In terms of raw score cut, the closest raw scores that would have provided the similar percentage in each achievement level were identified and presented in Table 21.

Table 19. The Post-Policy Measurement Review Panel Results for the Mean Recommended Percent of Students That Should Be Classified as Level 2 and Above or Level 3, Round 1

	Level 2 and Above	Level 3
Mean	72.6	26.1
Median	72.5	25.0
Maximum	95.0	50.0
Minimum	45.0	15.0
Standard Deviation	10.7	6.5

Table 20. The Post-Policy Measurement Review Panel Results for the Mean Recommended Percent of Students That Should Be Classified as Level 2 and Above or Level 3, Final Round

	Level 2 and Above	Level 3
Mean	73.6	24.8
Median	73.5	25.0
Maximum	95.0	40.0
Minimum	55.0	15.0
Standard Deviation	8.9	4.7

Table 21. Corresponding Raw Score Cuts for the Two Rounds Based on Impact Data Results

	Round 1	Final Round
Level 2	39	38
Level 3	65	65

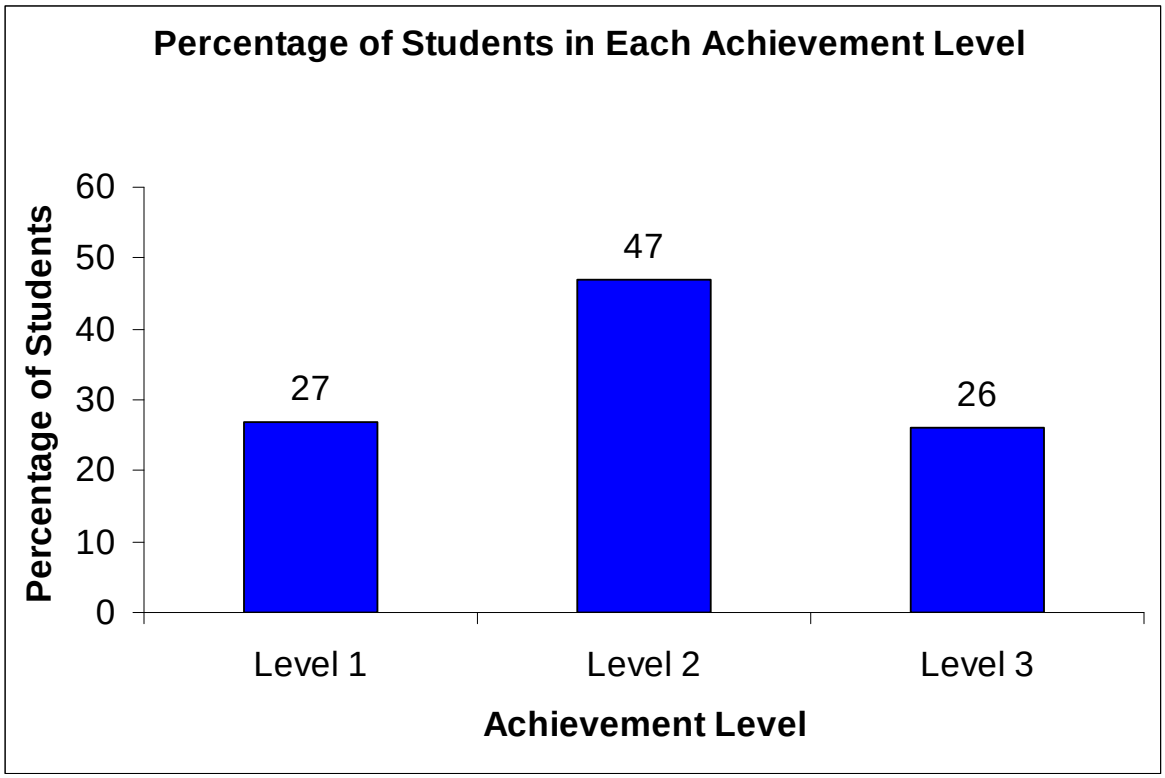

Figure 5. The Percentage of Students in Each Achievement Level Based on the Mean Recommendations from the Post-Policy Measurement Review Panel, Round 1

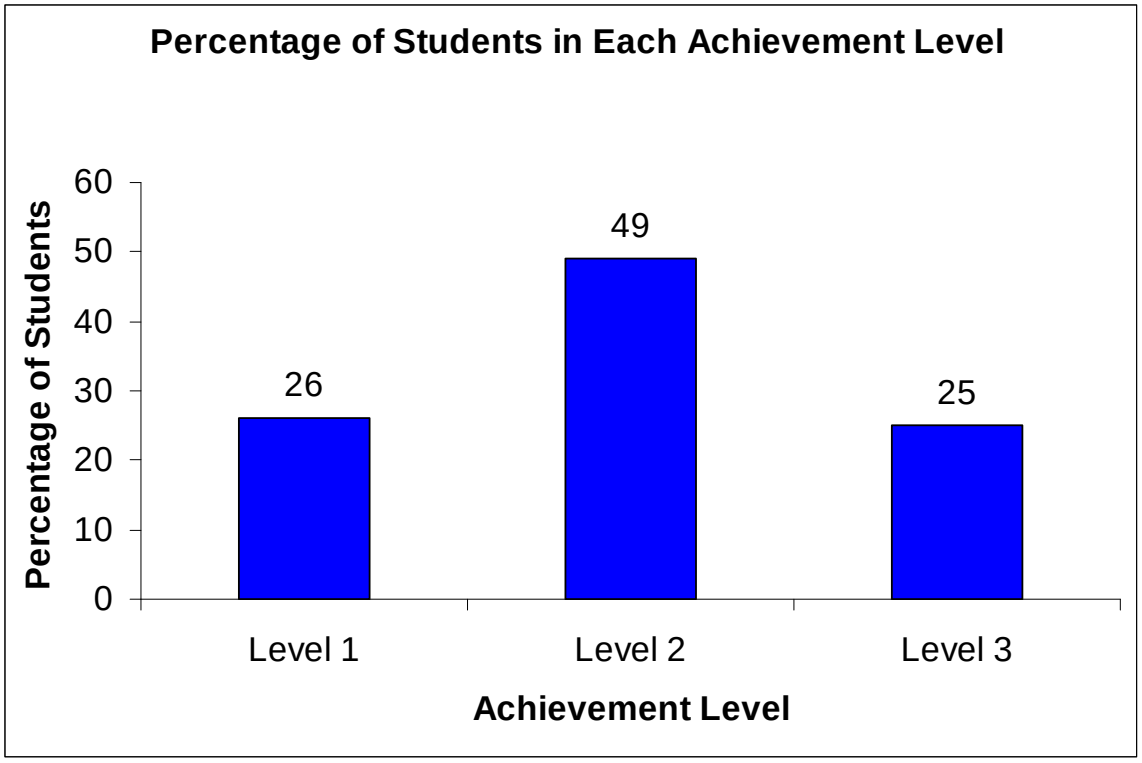

Figure 6. The Percentage of Students in Each Achievement Level Based on the Mean Recommendations from the Post-Policy Measurement Review Panel, Final Round

Evaluations

An exit survey was completed by each panelist following the completion of the Post-Policy Measurement Review Panel meeting. Panelists answered each question, using a scale of 1–5, 1 being "**totally disagree**" and 5 being "**totally agree**." The survey questions and the results are shown, below, in Table .

Table 22. Questionnaire Results for the Post-Policy Measurement Review Panel

Question	Mean	Median	Minimum	Maximum	Standard Deviation
The method for making recommendations on the ideal percentage of students who should be classified in each achievement level was conceptually clear.	4.7	5.0	2.0	5.0	0.6
I had a good understanding of the results from the earlier meeting of Pre-Policy Measurement Review.	4.9	5.0	4.0	5.0	0.3
I had a good understanding of the results from the earlier Item Mapping Meeting.	4.8	5.0	4.0	5.0	0.4
After the first round of ratings, I felt comfortable with the method for making recommendations.	4.6	5.0	2.0	5.0	0.7
After the second round of ratings, I felt comfortable with the method for making recommendations.	4.8	5.0	2.0	5.0	0.6
I found the feedback on the recommendations of other panelists useful in making my second round recommendations.	4.7	5.0	4.0	5.0	0.5
I found the feedback on the overall group recommendation useful in making my second round recommendations.	4.7	5.0	4.0	5.0	0.5
I feel confident that the final cut score recommendations reflect the achievement levels associated with the New York State Regents Examination in Algebra 2/Trigonometry.	4.4	5.0	3.0	5.0	0.7

Final Recommendation and Decision

As described in previous sections, NYSED conducted a formal standard setting process with Pearson that consisted of the following activities:

1) Pre-Policy Measurement Review
2) Item Mapping Standard Setting
3) Post-Policy Measurement Review

The three activities went according to plan and reflected both TAG's overview and recommended suggestions. The standard setting groups were diverse and representative of New York State. All groups adhered to instructions and processes put forward to them from the lead standard setting staff of Pearson. All activities were formally observed by the Office of State Assessment's Senior Managers and psychometric research staff.

After the standard setting activities, a conference call was set up between NYSED management and research staff, TAG members and Pearson psychometricians leading the standard setting meetings. The standard setting process and results were presented to TAG, and TAG formally endorsed NYSED's administration of the standard setting processes.

Here is a summary of the final standard setting recommendations from the content perspective and the policy perspective:

- Content perspective:
  - Raw score cut for a scale score of 65 was 46
  - Raw score cut for a scale score of 85 was 65
- Policy perspective:
  - Final percentage of students reaching Level 2 and above (with a scale score of 65 or higher) should be 73.6%, which, using impact data, translated to a raw score of 38
  - Final percentage of students reaching Level 3 (with a scale score of 85 or higher) should be 24.8%, which, using impact data, translated to a raw score of 65

After careful considerations of factors such as the nature of the assessment, how rigorous the new curriculum is and how teachers in the field are adjusting to teach it, the role of the assessment in students' learning and advance in high school, its desired impact, and so on, the senior management team made recommendations on the cut scores to the New York State Commissioner of Education. The Commissioner decided on the final cut scores for the Regents Examination in Algebra 2/Trigonometry and they are presented below:

- The final raw score cut for a scale score of 65 was 46, the same as the final recommended cut from the item mapping final recommendation.

- The final raw score cut for a scale score of 85 was 65, the same as the final recommended cut from the item mapping final recommendation.

The final impact of the cut scores on the students taking the Algebra 2/ Trigonometry assessment in June 2010 is as follows:

- 0–64 (Level 1), 39.1%
- 65–84 (Level 2), 35.2%
- 85–100 (Level 3), 25.7%

As the report described, the standard setting process has been conducted carefully and best psychometric practices have been followed. The policy decisions adhered to sound measurement principles to guarantee a thoughtful setting of cut scores, and NYSED is staying consistent with the approaches that have been integrated to the state's standard setting processes and have been used with the Grades 3–8 Testing Program and the Regents Examination in Integrated Algebra and the Regents Examination in Geometry.

Appendix A

Agenda for the Pre-Policy Review Panel Meeting

University of the State of New York
State Education Department

PEARSON

Agenda for Pre-Policy Measurement Review Meeting
Regents Examination in Algebra 2/Trigonometry
April 15, 2010

Registration	7:30-8:00
Welcome and Introductions	8:00-8:30
Purpose of Meeting	
Introductions	
Review of Agenda	
Non-disclosure Agreement Form	
Reimbursement Forms	
Opening Remarks	8:30-9:00
David Abrams, Assistant Commissioner	
Overview of Measurement Review	9:00-10:00
Purpose	
Methodology	
BREAK	10:00-10:15
Presentation of Related Assessments	10:15-11:00
New York State Grade 8 Mathematics Test	
Regents Examination in Mathematics A	
Regents Examination in Mathematics B	
Regents Examination in Integrated Algebra	
Regents Examination in Geometry	
National Assessment of Education Progress (NAEP)	
Break into Pre-assigned Groups	11:00
Discuss Assessment Data and Guiding Questions	11:00-11:45
Round 1 Recommendations	11:45-12:00
Readiness Form	
Review Method	
Collect Recommendations	
LUNCH	12:00-1:00
Large Group Feedback	1:00-2:00
Break into Pre-assigned Groups	2:00
Discuss Feedback in Small Groups	2:00-2:30
Round 2 Recommendations	2:30-2:45
Readiness Form	
Review Method	
Collect Recommendations	
BREAK	2:45-3:00
Feedback on Final Recommendations	3:00-3:15
Complete Survey	3:15-3:30
End of Day Activities (Check in Materials)	3:30-3:45

Appendix B

Recommendation Form for the Pre-Policy Review Panel Meeting

Rating Form

REGENTS EXAMINATION IN ALGEBRA 2/TRIGONOMETRY
PRE-POLICY MEASUREMENT REVIEW
RATING FORM

Directions: Your task is to recommend the percentage of students who should be classified in each achievement level on the Regents Examination in Algebra 2/Trigonometry. First, recommend the percentage of students who should be classified as *Level 2 and above*. Next, recommend the percentage of students who should be classified as *Level 3*. Make your recommendations using values that are multiples of 5, for example, 50, 55 or 60.

Panelist ID: _____

Room: _____

Round 1

Lowest Acceptable Percentage	
% Level 2 & Above	% Level 3
Highest Acceptable Percentage	
% Level 2 & Above	% Level 3

Round 2

Lowest Acceptable Percentage	
% Level 2 & Above	% Level 3
Highest Acceptable Percentage	
% Level 2 & Above	% Level 3

Appendix C

Demographic Questionnaire for the Item Mapping Committee Meetings

New York State Regents Examination in Algebra 2/Trigonometry
Pre-Policy Measurement Review
Panelist Information Sheet

Panelist ID: _____

Please provide the following demographic information. This information will be used to describe the general characteristics of the panelists who are recommending standards as members of the Pre-Policy Measurement Review Committee.

Your Current Position:

_____

_____

_____

Gender (circle one): Male Female

Ethnicity:

_____

Years of Work Experience in Education:

_____

Experience with Special Population (Circle all that apply):

Special Education ELL

Compared to other school districts in New York State, how would you describe the size of your district (circle one)?

Large Medium Small

Compared to other school districts in New York State, how would you describe the location of your district (circle one)?

Urban Suburban Rural

Compared to other school districts in New York State, how would you describe the geographic location of your district (circle one)?

North Country Long Island NYC Lower and Mid-Hudson

Capital Region Central NY Western NY

Appendix D

Agenda for the Item Mapping Committee Meeting

University of the State of New York
State Education Department

David Abrams
Assistant Commissioner
Office for Standards, Assessments and Reporting

PEARSON

2510 North Dodge Street
Iowa City, IA 52245
(800) 627-7990

Recommendations for Setting Achievement Levels for the New York State Regents Examination in Algebra 2/Trigonometry Agenda

DAY 1—June 21, 2010

Registration	7:30–8:00
Welcome and Overview—David Abrams Examination Development Overview of Standard Setting Process	8:00–8:45
Opening Remarks—Paul Nichols Introduction Why are you here? Agenda Non-Disclosure Forms Reimbursement Forms	8:45–9:15
Standard Setting Process—Paul Nichols Purpose Item Mapping Methodology	9:15–9:45
Overview of the Regents Examination in Algebra 2/Trigonometry History Purposes Test Specifications	9:45–10:00
BREAK	10:00–10:15
Introduce Achievement Level Descriptors	10:15–10:30
Construct Achievement Level Descriptors Small Group Discussion	10:30–12:00
LUNCH Train Table Leaders	12:00–1:00

Construct Achievement Level Descriptors Large Group Discussion	1:00–1:30
Break to Two Committee Rooms Introduction	1:30–1:45
Complete Selected Algebra 2/Trigonometry Items on the Exam	1:45–2:45
BREAK Assign Panelist IDs	2:45–3:00
Overview of Standard Setting Item Mapping Ordered Item Booklet Item Map Ratings Forms	3:00–3:30
Practice Round	3:30–3:45
Round 1 Standard Setting Readiness Form Review Method Collect Completed Rating Forms	3:45–4:45
End of Day Activities Review Day 2 Schedule Check in Materials	4:45–5:00

END OF DAY 1

DAY 2—June 22, 2010

Registration	8:00–8:15
Review Schedule, Answer Questions (both committees)	8:15–8:30
Reconvene in Committee Rooms	8:30
Feedback Small Group Discussion of Table Agreement Data Small Group Discussion of Impact Data	8:30–9:15
Round 2 Ratings Readiness Form Review Method Collect Completed Rating Forms	9:15–10:15
BREAK	10:15–10:45
Feedback Small Group Discussion of Table Agreement Data Committee Discussion of Group Agreement Data Committee Discussion of Impact Data	10:45–11:30
Round 3 Ratings Readiness Form Review Method Collect Completed Rating Forms	11:30–12:00
LUNCH	12:00–1:00
Feedback	1:00–1:15
Revisit Achievement Level Descriptors Revise Achievement Level Descriptors	1:15–2:00
Closing Remarks—David Abrams	2:00–2:15
Complete Survey	2:15–2:30
End of Day Activities Check in Materials	2:30–3:00

University of the State of New York
State Education Department

David Abrams
Assistant Commissioner
Office for Standards, Assessments and Reporting

PEARSON

2510 North Dodge Street
Iowa City, IA 52245
(800) 627-7990

Agenda for Standard Setting Synthesis Group Discussion
New York State Regents Examination in Algebra 2/Trigonometry
June 22, 2010

Overview of the Synthesis Review Purpose Methodology	3:00–3:10
Presentation from Each Committee Presentation from Committee A Presentation from Committee B	3:10–3:30
Group Discussion Focus on Items Around the Cuts Focus on Skills and Knowledge	3:30–4:30
Rating	4:30–5:00

Appendix E

Achievement Level Descriptors

New York State
Regents Examination in Algebra 2/Trigonometry
Achievement Levels

■ **Not Passing** – A ***not passing*** student is unable to demonstrate, on demand, proficiency in understanding the content and concepts required for commencement-level achievement in any or most of the learning standards and key ideas assessed. – A ***not passing*** student is unable to demonstrate on demand, proficiency in the skills required for commencement-level achievement in any or most of the learning standards and key ideas assessed. – A ***not passing*** student is unable to demonstrate, on demand, evidence of an ability to apply the content, concepts, and skills required to meet any or most of the demands of productive adult citizenship, the workplace, and postsecondary education.
■ **Passing** – A ***passing*** student is able to demonstrate, on demand, knowledge of the content and concepts required for commencement-level achievement in each of the learning standards and key ideas assessed. – A ***passing*** student is able to demonstrate, on demand, the skills required for commencement-level achievement in each of the learning standards and key ideas assessed. – A ***passing*** student is able to apply, on demand, the content, concepts, and skills required to meet the demands of productive adult citizenship, the workplace, and postsecondary education.
■ **Passing with Distinction** – A ***passing with distinction*** student is able to demonstrate, on demand, evidence of superior understanding of the content and concepts required for commencement-level achievement in each of the learning standards and key ideas assessed. – A ***passing with distinction*** student is able to demonstrate, on demand, evidence of superior skills required for commencement-level achievement in each of the learning standards and key ideas assessed. – A ***passing with distinction*** student is able to demonstrate, on demand, evidence of superior ability to apply the content, concepts, and skills required to meet the demands of productive adult citizenship, the workplace, and postsecondary education.

New York State Regents Examination in Algebra 2/Trigonometry
Learning Standard

Mathematics, Science, and Technology—Standard 3

Students will:

- understand the concepts of and become proficient with the skills of mathematics;
- communicate and reason mathematically;
- become problem solvers by using appropriate tools and strategies;
- through the integrated study of number sense and operations, algebra, geometry, measurement, and statistics and probability.

New York Regents Examination in Algebra 2/Trigonometry
Achievement Level Descriptors
June 21–22, 2010

Level 2 (65–84)

1. Proficient at using a calculator for standard type problems
2. Calculator dependent
3. Would realize the answer is wrong, but not know how to do it correctly
4. Basic algebra proficient
5. Proficient in math, but not enthusiastic about it
6. Can solve an equation, but struggles with how to get from a word problem to the equation
7. Limited math vocabulary
8. More limited to thinking about solving a problem. Unable to go outside of the box to solve how it was taught.
9. Can do multiple-choice
10. May not fully understand what the question is asking, but does show some ability to attempt
11. Some calculator knowledge
12. Working knowledge of algebra, can do the simple things
13. Successful with problems they are familiar with
14. Proficient with graphing calculator
15. Working knowledge of curriculum
16. Solve, simplify, and check answer
17. Have memorized most formulas necessary
18. Mechanical or computational errors, may do some guessing
19. Able to present logical justifications for an answer
20. Attempt most if not all open-ended questions
21. Math skills are not strong, but persistent in an attempt to solve
22. Find least common denominator and apply least common denominator
23. Attempts every question
24. Formula recognition
25. Adequate calculator ability
26. Understand reasonableness in an answer
27. May do one step of a problem, but not apply the answer to the next step
28. Have some math knowledge and show it by attempting every problem
29. Fragmented competence in math concepts
30. Firm grasp of vocabulary.
31. May use only method to solve a problem
32. Generally competent and able to visualize the problem
33. High level of concern for passing, perseverance
34. Make conceptual and computational errors
35. Unable to demonstrate application skills
36. Makes a valid attempt

Level 3 (85–100)

1. Creative and manipulative with a calculator, can check an answer wheter using a calculator or not
2. Motivated to do well and eager to do math
3. Attempt to figure out problems even if haven't seen it before
4. Can solve an equation and find different ways of checking it
5. Able to take problems and make equations out them easily
6. Strong vocabulary skills
7. Higher order level of problem solving and thinking
8. Check work, spend more time
9. Work really hard to get every point possible, apply what they can do in order to get some points
10. Strong calculator knowledge and can use it to figure things out even if not absolutely sure how to tackle the problem
11. Look at questions in a multifaceted way, with more than one way to approach the question
12. Understand the mathematical vocabulary at this level
13. Would see many different ways to solve problem
14. Confident in how to check their work and will check and correct mistakes
15. Their work is elegant at this level
16. They take something they know and work step by step from beginning to end
17. Even when have not seen the problem before, able to work through and tackle the problem
18. Simple computational errors, carelessness
19. May be holes in the work or in the test and they do not get it
20. Can do some of the higher level questions, but not all
21. Have complete control of calculator
22. Can make connections to geometry when appropriate
23. Thorough completion of work
24. Extensive math vocabulary
25. Excellent algebra skills
26. Proceed through a problem in a logical manner
27. Able to round correctly without double-rounding within the problem
28. Understand the different types of questions and can jump between different areas
29. Even if make mistakes, the answer is reasonable
30. Very confident, resourceful, and mathematically fluent
31. Can look at every problem algebraically
32. Demonstrates conceptual understanding with few computational errors
33. Critical thinking skills
34. Few careless errors

Level 1 (0–64)

1. Basic algebra skills, but make basic mistakes
2. Calculator dependent and do not recognize when an answer is wrong
3. Have difficulty recognizing the correct formula to use for a problem
4. Not motivated and do not expect to do well in the subject
5. Not proficient in algebra
6. Often apply the wrong formula
7. Substitutions done incorrectly
8. Can use the calculator, but not real proficient
9. Good with simple substitutions and repetition
10. Lots of basic algebraic errors
11. Many blanks in open-ended questions with inappropriate work when work is shown
12. Will randomly choose formula off of the reference sheet
13. Able to begin some of the problems
14. Some idea of the curriculum
15. Limited vocabulary
16. Cannot think abstractly
17. Several incoherent starts or blanks
18. Underestimates the difficulty of a question

New York State Regents Examination in Algebra 2/Trigonometry
Top Three Distinctive Features of Threshold Level Students

Just 65 Student

1. Persistent in attempting an answer , but may not finish
2. Adequate algebra skills
3. Moderate ability to use calculator for problem solving

Just 85 Student

1. Solid algebra skills
2. Flexible, analytical, able to use multiple methods to solve problems, able to think outside of the box, and a higher order thinker
3. Excellent calculator skills; will use a calculator as a tool, not a prop

Appendix F

Recommendation Form for the Item Mapping Committee Meetings

New York State Regents Examination in Algebra 2/Trigonometry
Item Position Recording Sheet

Panelist ID _____

		Round 1	Initials		Round 2	Initials		Round 3	Initials
65	Recommended Cut								
85	Recommended Cut								

Appendix G

Agenda for the Post-Policy Review Committee Meeting

University of the State of New York
State Education Department

David Abrams
Assistant Commissioner
Office for Standards, Assessments and Reporting

PEARSON

2510 N Dodge Street
Iowa City, IA 52245
(800) 627-7990

New York State Regents Examination in Algebra 2/Trigonometry
Post-Policy Measurement Review
Agenda

Registration	4:30–4:45
Dinner	4:45–5:30
Welcome and Overview—David Abrams Introductions Overview of Standard Setting Process	5:30–5:45
Overview of Post-Policy Measurement Review Logistics Purpose Methodology Present Results from Standard Setting Discussions	5:45–6:30
Round 1 Recommendations Readiness Form Review Method Collect Recommendation	6:30–6:45
Break	6:45–7:00
Discussions Round 1 Results Group Discussion	7:00–7:20
Round 2 Recommendations Readiness Form Review Method Collect Recommendation	7:20–7:30

Appendix H

Recommendation Form for the Post-Policy Review Committee Meeting

NEW YORK STATE
REGENTS EXAMINATION IN ALGEBRA 2/TRIGONOMETRY
POST-POLICY MEASUREMENT REVIEW
RATING FORM

Directions: Your task is to recommend the percentage of students who should be classified in each achievement level on the Regents Examination in Algebra 2/Trigonometry. First, recommend the percentage of students who should be classified as *Level 2 (65-84) and Above*. Next, recommend the percentage of students who should be classified as *Level 3 (85-100)*. Make your recommendations using values that are multiples of 5, for example, 50, 55, or 60.

Panelist ID: _____

Round 1

Lowest Acceptable Percentage	
% Level 2 & Above	% Level 3
Highest Acceptable Percentage	
% Level 2 & Above	% Level 3

Round 2

Lowest Acceptable Percentage	
% Level 2 & Above	% Level 3
Highest Acceptable Percentage	
% Level 2 & Above	% Level 3

Appendix I

Ordered Item Booklet

Table I1. Ordered Item Book

Page Number	Item ID	θ with RP 0.67	Page Number	Item ID	θ with RP 0.67
1	OPITEM_1	-2.277	51	ANCHOR2_9_3	0.722
2	ANCHOR2_4	-1.567	52	OPITEM_14	0.743
3	OPITEM_2	-1.327	53	OPITEM_17	0.743
4	OPITEM_33_1	-1.050	54	ANCHOR2_9_4	0.746
5	ANCHOR1_2	-0.877	55	OPITEM_39_5	0.748
6	OPITEM_5	-0.807	56	ANCHOR1_11_4	0.789
7	OPITEM_3	-0.567	57	OPITEM_35_2	0.792
8	OPITEM_6	-0.497	58	ANCHOR1_9_2	0.801
9	OPITEM_31_1	-0.298	59	ANCHOR1_4	0.803
10	OPITEM_4	-0.227	60	ANCHOR2_5	0.803
11	OPITEM_7	-0.197	61	ANCHOR1_5	0.813
12	OPITEM_8	-0.187	62	OPITEM_18	0.843
13	OPITEM_38_1	-0.150	63	OPITEM_30_2	0.852
14	ANCHOR1_3	-0.147	64	OPITEM_36_2	0.917
15	OPITEM_10	-0.047	65	OPITEM_19	0.933
16	ANCHOR1_11_1	-0.010	66	OPITEM_29_2	0.940
17	OPITEM_39_1	0.047	67	ANCHOR2_3	0.943
18	ANCHOR1_9_1	0.067	68	OPITEM_13	0.983
19	OPITEM_35_1	0.072	69	ANCHOR1_8_2	0.988
20	OPITEM_12	0.093	70	ANCHOR2_9_5	1.009
21	OPITEM_32_1	0.107	71	OPITEM_16	1.033
22	OPITEM_11	0.123	72	OPITEM_34_2	1.047
23	ANCHOR1_6	0.133	73	ANCHOR2_8_1	1.052
24	OPITEM_29_1	0.192	74	ANCHOR2_6	1.093
25	OPITEM_34_1	0.211	75	OPITEM_28_1	1.120
26	OPITEM_39_2	0.232	76	OPITEM_36_3	1.141
27	ANCHOR1_11_2	0.265	77	OPITEM_22	1.163
28	OPITEM_30_1	0.278	78	OPITEM_38_4	1.169
29	OPITEM_38_2	0.279	79	ANCHOR1_10_3	1.177
30	OPITEM_15	0.313	80	OPITEM_25	1.283
31	ANCHOR2_2	0.333	81	OPITEM_23	1.283
32	ANCHOR1_7	0.363	82	ANCHOR1_10_4	1.310
33	ANCHOR2_1	0.363	83	ANCHOR2_10_2	1.344
34	OPITEM_39_3	0.408	84	OPITEM_39_6	1.369
35	ANCHOR1_10_1	0.468	85	OPITEM_37_1	1.408
36	OPITEM_26	0.473	86	OPITEM_32_2	1.439
37	ANCHOR2_9_1	0.478	87	ANCHOR2_9_6	1.599
38	ANCHOR2_7	0.533	88	ANCHOR2_11_2	1.642
39	ANCHOR1_11_3	0.544	89	OPITEM_36_4	1.656

Page Number	Item ID	θ with RP 0.67	Page Number	Item ID	θ with RP 0.67
40	ANCHOR1_8_1	0.549	90	OPITEM_27	1.783
41	ANCHOR1_1	0.553	91	OPITEM_37_2	1.788
42	OPITEM_39_4	0.557	92	OPITEM_33_2	1.805
43	ANCHOR1_10_2	0.583	93	ANCHOR2_8_2	1.828
44	OPITEM_9	0.593	94	OPITEM_21	1.913
45	ANCHOR2_9_2	0.608	95	OPITEM_28_2	2.084
46	ANCHOR2_11_1	0.610	96	OPITEM_24	2.113
47	ANCHOR2_10_1	0.652	97	OPITEM_37_3	2.676
48	OPITEM_36_1	0.680	98	OPITEM_31_2	3.179
49	OPITEM_20	0.683	99	OPITEM_37_4	3.362
50	OPITEM_38_3	0.709			

Table I2. Raw to θ Conversion Table

Raw Score	θ	CSEM	Raw Score	θ	CSEM
0	-5.638	1.840	45	0.172	0.206
1	-4.896	1.025	46	0.214	0.205
2	-4.154	0.740	47	0.256	0.205
3	-3.702	0.615	48	0.298	0.205
4	-3.370	0.541	49	0.340	0.205
5	-3.105	0.491	50	0.382	0.205
6	-2.882	0.454	51	0.424	0.206
7	-2.689	0.425	52	0.467	0.207
8	-2.518	0.402	53	0.510	0.208
9	-2.364	0.383	54	0.553	0.209
10	-2.224	0.367	55	0.597	0.211
11	-2.094	0.353	56	0.642	0.213
12	-1.974	0.341	57	0.688	0.215
13	-1.862	0.330	58	0.735	0.217
14	-1.756	0.321	59	0.783	0.220
15	-1.655	0.312	60	0.832	0.223
16	-1.560	0.305	61	0.882	0.226
17	-1.470	0.298	62	0.934	0.229
18	-1.383	0.291	63	0.987	0.233
19	-1.300	0.285	64	1.042	0.237
20	-1.220	0.280	65	1.099	0.241
21	-1.143	0.275	66	1.158	0.245
22	-1.069	0.270	67	1.220	0.250
23	-0.998	0.265	68	1.284	0.255
24	-0.928	0.261	69	1.350	0.261
25	-0.861	0.257	70	1.420	0.267
26	-0.796	0.253	71	1.493	0.274
27	-0.733	0.250	72	1.570	0.281
28	-0.672	0.246	73	1.651	0.289
29	-0.612	0.243	74	1.737	0.297
30	-0.554	0.239	75	1.828	0.306
31	-0.497	0.236	76	1.925	0.317
32	-0.442	0.233	77	2.029	0.328
33	-0.389	0.230	78	2.141	0.342
34	-0.336	0.227	79	2.263	0.357
35	-0.285	0.225	80	2.397	0.375
36	-0.235	0.222	81	2.546	0.398
37	-0.187	0.219	82	2.715	0.425
38	-0.139	0.217	83	2.911	0.461
39	-0.092	0.215	84	3.146	0.511
40	-0.047	0.213	85	3.444	0.586
41	-0.002	0.211	86	3.858	0.712
42	0.042	0.209	87	4.558	1.003
43	0.086	0.208	88	5.258	1.827
44	0.129	0.207			