New York State Regents Examination in Integrated Algebra

Standard Setting Technical Report

Prepared by:
Pearson
November 25, 2008

Table of Contents

Executive Summary **5**

Pre-Policy Measurement Review Panel 5
Item Mapping Standard Setting 6
Post-Policy Measurement Review Panel 8

Technical Report **10**

Pre-Policy Measurement Review Panel 10
Panelists *10*
Method and Procedure *11*
Results *13*
Item Mapping Standard Setting 17
Panelists *17*
Method *19*
Procedure *26*
Results *29*
Panelist Variability *34*
Evaluations *38*
Post-Policy Measurement Review Panel 40
Panelists *40*
Method and Procedure *40*
Results *41*
Evaluations *43*

Appendix A **44**

Appendix B **47**

Appendix C **49**

Appendix D **51**

Appendix E **55**

Appendix F **57**

Appendix G **60**

Appendix H **62**

List of Tables

Table 1. Numbers of Male and Female Panelists on the Committee 10
Table 2. Summary of the Ethnic Representation of the Panelists on the Committee 11
Table 3. Round One Results for the Recommended Percentage of Students Who Should Be Classified as Level 2 and Above or Level 3 13
Table 4. Round Two Results for the Recommended Percentage of Students Who Should Be Classified as Level 2 and Above or Level 3 14
Table 5. Questionnaire Results for the Pre-Policy Measurement Review Panel Meeting (N=21) 16
Table 6. Numbers of Male and Female Panelists on Committee A 17
Table 7. Summary of the Ethnic Representation of the Panelists on Committee A 17
Table 8. Summary of Years of Educational Experience for the Panelists on Committee A 17
Table 9. Summary of District Size for the Panelists on Committee A 18
Table 10. Summary of District Location for the Panelists on Committee A 18
Table 11. Numbers of Male and Female Panelists on Committee B 18
Table 12. Summary of the Ethnic Representation of the Panelists on Committee B 18
Table 13. Summary of Years of Educational Experience of the Panelists on Committee B 19
Table 14. Summary of District Size for the Panelists on Committee B 19
Table 15. Summary of District Location for the Panelists in Committee B 19
Table 16. Committee A Cut Scores for Achievement Levels 2 and 3 29
Table 17. Committee B Cut Scores for Achievement Levels 2 and 3 31
Table 18. Cut Scores Across Both Committees for Achievement Levels 2 and 3 33
Table 19. The G-Theory Study for the Committee A Level 2 Cut Score 35
Table 20. G-Theory Study for the Committee A Level 3 Cut Score 36
Table 21. G-Theory Study for the Committee B Level 2 Cut Score 36
Table 22. G-Theory Study for the Committee B Level 3 Cut Score 37
Table 23. Questionnaire Results for Committee A 38
Table 24. Questionnaire Results for Committee B 39
Table 25. Post-Policy Measurement Review Panel Results for the Mean Recommended Percentage of Students Classified as Level 2 and Above or Level 3 41
Table 26. Questionnaire Results for the Post-Policy Measurement Review Panel 43
Table H 1. Ordered Item Book. 62
Table H 2. Raw to θ Conversion Table. 63

List of Figures

Figure 1. **Mean Percentage of Students in Each Achievement Level Following Round Two** **15**
Figure 2. **Percentage of Students in Each Achievement Level, Using Committee A Cut Scores After the Round Three Final Rating** **30**
Figure 3. **Percentage of Students in Each Achievement Level, Using Committee B Cut Scores After the Round Three Final Rating** **32**
Figure 4. **Percentage of Students in Each Achievement Level, Using Cut Scores Across Both Committees After the Round Three Final Rating** **34**
Figure 5. **Mean Percentage of Students in Each Achievement Level, Using the Mean Recommendations from the Post-Policy Measurement Review Panel** **42**

Executive Summary

This executive summary briefly describes the three sets of standard setting activities for the Regents Examination in Integrated Algebra.

- Pre-Policy Measurement Review Panel
- Item Mapping Standard Setting
- Post-Policy Measurement Review Panel

On Thursday, April 17, 2008, Dr. Paul Nichols and Dr. Ye Tong, both of Pearson, conducted the Pre-Policy Measurement Review Panel meeting in Albany, New York. This meeting was convened to provide recommendations for the ideal percentage of New York students who should be classified in each achievement level on the Regents Examination in Integrated Algebra. On Monday, June 23, and Tuesday, June 24, Dr. Paul Nichols and Dr. Ye Tong conducted an item mapping procedure for two committees to recommend achievement standards. Finally, on the afternoon of Tuesday, June 24, Dr. Paul Nichols conducted the Post-Policy Measurement Review Panel meeting. This meeting was convened with panelists from the Pre-Policy Measurement Review Panel to integrate results from the Pre-Policy Measurement Review Panel meeting and the two committees from the item mapping meeting. Both the item mapping meeting and the Post-Policy Measurement Review Panel meeting were held in Albany. The outcomes of the final round of activities are described in this executive summary. A more detailed description, including results from each round, will be provided in the subsequent Standard Setting Technical Report.

Pre-Policy Measurement Review Panel

Achievement Levels	Scale Scores
Level 1	0 – 64
Level 2	65 – 84
Level 3	85 – 100

A total of 21 panelists attended this meeting. All panelists provided voluntary demographic information. Complete demographic information for the panelists is summarized in the body of the technical report

Panelists participated in two rounds of discussion in which they were asked to make individual “high” and “low” recommendations as to the ideal percentage of students who should be classified in each achievement level. For example, a

committee member could recommend that 25%–30% of students be classified as Level 3 on the Regents Examination in Integrated Algebra.

The meeting began with introductions of the facilitators and New York State Education Department (NYSED) staff. 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 impact data from a number of New York State examinations and from the National Assessment of Educational Progress (NAEP). Panelists then broke into two groups, one with 10 panelists and the other with 11. The groups met in separate rooms, where they 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. Panelists then again broke into two groups and completed a second round of recommendations. For those panelists whose schedules permitted, the final average recommendations were presented without further discussion.

The mean recommended percentage of students was computed by averaging both “low” and “high” recommendations across all panelists. For round two, the mean of the panelists’ recommendations for the ideal percentage of students who should be classified as Level 2 and above was 82 percent and for the ideal percentage of students who should be classified as Level 3 was 25 percent.

Exit surveys were administered following the completion of the Pre-Policy Measurement Review Panel meeting.

Item Mapping Standard Setting

Two committees of New York educators were convened June 23 and June 24, 2008, in Albany to recommend standards for the Regents Examination in Integrated Algebra. The first, Committee A, had 31 educators, and the second, Committee B, had 30 educators.

Complete demographic information for the panelists is summarized in the body of the technical report.

The morning was devoted to introductions of the staff, a description of standard setting, and a description of the Regents Examination in Integrated Algebra.

Following the midmorning break, the two committees met in their separate conference rooms. Each committee began the process of reviewing achievement level descriptors. This process required several hours, and resulted in a set of descriptors for each achievement level (Level 1, Level 2, and Level 3).

At approximately 2:30 p.m. on Monday, June 23, each committee began the standard setting process. The item mapping procedure was the methodology used. In this procedure, panelists are asked to identify the item in an ordered item book that is the last item that a threshold student at a given level would be able to correctly answer. 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 2008 administration and anchor items from a field-test administration from 2007. Items were sorted from least to most difficult, using their Rasch item difficulty values.

The standard setting process consisted of three rounds of judgments. 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 five or six. They were provided the page number for each panelist on the basis of the round one ratings in addition to the mean, median, minimum, and maximum page number for each table at each level for that table. Panelists were informed that there was no intention that they come to consensus on their page number placement, but they should discuss differences to get a feel for why differences exist.

In addition, panelists were provided a list of item *p*-values. Finally, panelists were provided a graphical display of the percentage of students in each achievement level, on the basis of the median cut scores from round one.[1] For both the *p*-values and the graphical display, a representative sample of approximately 144,000 students was used from the June 2008 operational administration of the Regents Examination in Integrated Algebra. Panelists were given time to discuss the appropriateness of the committee level cut scores given the proportion of students who would fall in each level.

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

Next, panelists were provided a graphical display of the percentage of students in each achievement level, on the basis of the median cut score from round two. A representative sample of approximately 144,000 students was used from the June 2008 operational administration of the Regents Examination in Integrated Algebra. Panelists were also provided with a graphical display of the percentage of students in each achievement level disaggregated for Grade 8 students and Grade 9 and above students.

Panelists were given time to discuss the appropriateness of the committee level cut scores given, the proportion of students who would fall in each level.

[1] The page number was translated into cut scores on the raw score scale. A detailed discussion is included in the body of the report.

Finally, panelists were asked to vote by a show of hands whether they favored immediately implementing the cut scores recommended or possibly phasing them in over several years.

Following round three, Committee A recommended a raw score of 40 as the cut score for Level 2 and a raw score of 65 as the cut score for Level 3. These cut scores would result in 43 percent of students classified as Level 1, 38 percent of students classified as Level 2, and 19 percent of students classified as Level 3 based on an operational sample of students collected in June 2008.

Committee B recommended a raw score of 48 as the cut score for Level 2 and a raw score of 65 as the cut score for Level 3. For Committee B, these cut scores would result in 57 percent of students classified as Level 1, 24 percent of students classified as Level 2, and 19 percent of students classified as Level 3 based on an operational sample of students collected in June 2008.

When the panelists were asked whether they favored immediately implementing the cut scores or possibly phasing them in over several years, two-thirds of the panelists voted for phasing them in over several years.

Exit surveys were administered following the completion of standard setting. The exit survey results are summarized in the body of the technical report. Additional analyses to assess judge variability and classification accuracy were performed on the standard setting data following the standard setting meeting. The results are summarized in the body of the report.

Post-Policy Measurement Review Panel

The Post-Policy Measurement Review Panel met on the afternoon of Tuesday, June 24, following the completion of the item mapping committee meetings. This meeting was convened to integrate results from the Pre-Policy Measurement Review Panel meeting and the two committees from the item mapping meeting.

This meeting was convened with 16 panelists from the Pre-Policy Measurement Review Panel. The Post-Policy Measurement Review Panel meeting began with introductions of the facilitators and NYSED staff. Panelists were then told 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. The product of this activity would be final recommendations for the percentage of students in each achievement level that reflected the influence of both meetings.

Following these initial activities, panelists reviewed results from the Pre-Policy Measurement Review Panel. They then were given an explanation of the item mapping methodology. They then reviewed the results for Committee A and Committee B from the item mapping meeting.

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 Pre-Policy Measurement Review Panel and both Committee A and Committee B from the item mapping meeting. They were then asked, as a committee, to verbally make final adjustments to the percentage of students in each achievement level.

Finally, panelists were asked to vote by a show of hands whether they favored immediately implementing the cut scores or possibly phasing them in over years.

Based on the panelists' independent recommendations, the mean of the panelists' recommendations for the ideal percentage of students who should be classified as Level 2 and above was 74 percent and for the ideal percentage of students who should be classified as Level 3 was 21 percent.

When asked to verbally make final adjustments to the percentage of students in each achievement level, the majority of panelists indicated they favored decreasing the percentage of students at Level 1 by five to ten percentage points and increasing the percentage of students at Level 2 and above by five to ten percentage points.

When the panelists were asked whether they favored immediately implementing the cut scores or possibly phasing them in over years, the panelists unanimously voted for phasing them in over several years. The panelists favored phasing in the raw score cut of 40 over a period of several years.

Technical Report

The standard setting process for the Regents Examination in Integrated Algebra consisted of three activities: the Pre-Policy Measurement Review Panel meeting, an item mapping procedure, 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 Regents Examination in Integrated Algebra. Students could be classified into the following three achievement levels on the Regents Examination in Integrated Algebra: the lowest level, Level 1 (0–64); Level 2 (65–84); and the highest level, Level 3 (85–100). Activities and results for the Pre-Policy Measurement Review Panel meeting, the item mapping procedure, and the Post-Policy Measurement Review Panel will be summarized separately in the following sections.

Pre-Policy Measurement Review Panel

On Thursday, April 17, 2008, Dr. Paul Nichols and Dr. Ye Tong, both of Pearson, conducted the Pre-Policy Measurement Review Panel meeting in Albany. This meeting was convened to provide recommendations for the ideal percentage of New York students who should be classified into each achievement level on the Regents Examination in Integrated Algebra. During this meeting, panelists participated in two rounds of discussions in which they were asked to make individual “high” and “low” recommendations as to the ideal percentage of students who should be classified into each achievement level. For example, a committee member could recommend that 25%–30% of students be classified as Level 3 on the Regents Examination in Integrated Algebra.

Panelists

A total of 21 panelists attended. All panelists provided voluntary demographic information. A summary of the numbers of male and female panelists is shown in Table 1, below.

Table 1. Numbers of Male and Female Panelists on the Committee

Male	Female
11	10

A summary of the ethnic representation of the panel is provided in Table 2, below.

Table 2. Summary of the Ethnic Representation of the Panelists on the Committee

Caucasian	Hispanic	African-American	Other
18	0	1	2

Method and Procedure

The Pre-Policy Measurement Review Panel meeting was held on Thursday, April 17. The agenda for the meeting is shown in Appendix A. The meeting began with a review of the agenda and introductions of the facilitators and NYSED staff. A description of standard setting was then presented, and panelists were introduced to the purpose of the meeting and the role they played in the process. Next, Pearson facilitators described the procedure that would be used for the meeting. Panelists were told that the outcome of this meeting would be their recommendations for an acceptable range of impact data for the Regents Examination in Integrated Algebra.

Following the mid-morning break, panelists were presented with a set of five questions. Panelists were instructed that these questions should guide their consideration of the test data they would subsequently see. The five questions were the following:

1. What type of differences do the participants expect between 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 would be the ideal percentage of students in each achievement level?
   - What variations from these ideal 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 between the current and new testing programs are acceptable?

4. What type of consistency does the panel expect between Grade 8 and Integrated Algebra?
   - What are the differences between the two testing programs?
   - Should the percentage of students in each achievement level be similar, even though the tests measure different knowledge and skills?
   - What differences between the Grade 8 and Integrated Algebra results are acceptable?
5. What, if any, consistency is expected between national data and New York State?
   - Should the percentage of students in each achievement level be similar, even though the testing programs are not similar?
   - What are the differences between New York State's testing program and NAEP?
   - What differences between the results for New York State's testing program and the NAEP testing program are acceptable?

Following presentation of these guiding questions, panelists reviewed a description and impact data for a number of New York State examinations and from the National Assessment of Educational Progress. The panelists reviewed the results of the following assessments:

1. New York State Grade 6 Mathematics Test
2. New York State Grade 7 Mathematics Test
3. New York State Grade 8 Mathematics Test
4. New York State Regents Examination in Mathematics A
5. New York State Regents Examination in Mathematics B
6. National Assessment of Educational Progress (NAEP) for New York State
7. National Assessment of Educational Progress (NAEP) across all states

Panelists then broke into two groups, one with 10 panelists and the other with 11. The groups met in separate rooms, where they first discussed reactions to the data just presented. They discussed each assessment in turn and how each is related to the Regents Examination in Integrated Algebra.

Following this discussion, panelists were instructed to recommend the ideal percentage of students who should be in each achievement level, using the ratings sheet shown in Appendix B. Panelists were instructed to think about the

design of the Regents Examination in Integrated Algebra, the design of the other assessments, the data that was presented, and the discussion around the guiding questions. Panelists made both a “high” and a “low” recommendation. They completed the first round of recommendations before breaking for lunch.

Following lunch, the committee reconvened to review the results from the first round. The committee was shown the minimum, maximum, mean, and median for each group and for the committee. A spokesperson from each group then summarized the discussion that had occurred in that group before the round one recommendations. Panelists were allowed to ask questions of other panelists and of the NYSED staff.

Then, panelists again broke into two groups and completed round two recommendations. For those panelists whose schedules permitted, the final average recommendations were presented without further discussion.

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. Table 3, below, summarizes, for round one, the panelists’ recommendations for the ideal percentage of students who should be classified as Level 2 and above and the ideal percentage of students who should be classified as Level 3.

Table 3. Round One Results for the Recommended Percentage of Students Who Should Be Classified as Level 2 and Above or Level 3

		Level 2 & Above	Level 3
Group 1	Mean	83.00	20.50
	Median	85.00	20.00
	Maximum	90.00	30.00
	Minimum	75.00	10.00
Group 2	Mean	74.09	28.64
	Median	75.00	22.50
	Maximum	90.00	75.00
	Minimum	50.00	10.00
Committee	Mean	78.33	24.76
	Median	80.00	20.00
	Maximum	90.00	75.00
	Minimum	50.00	10.00

Table 4, below, summarizes, for round two, the panelists' recommendations for the ideal percentage of students who should be classified as Level 2 and above and the ideal percentage of students who should be classified as Level 3.

Table 4. Round Two Results for the Recommended Percentage of Students Who Should Be Classified as Level 2 and Above or Level 3

		Level 2 & Above	Level 3
Group 1	Mean	83.75	20.50
	Median	85.00	20.00
	Maximum	90.00	30.00
	Minimum	80.00	15.00
Group 2	Mean	80.45	29.09
	Median	80.00	27.50
	Maximum	90.00	55.00
	Minimum	70.00	15.00
Committee	Mean	82.02	25.00
	Median	80.00	25.00
	Maximum	90.00	55.00
	Minimum	70.00	15.00

Figure 1 shows the percentage of students in each achievement level, using the mean recommendations from round two.

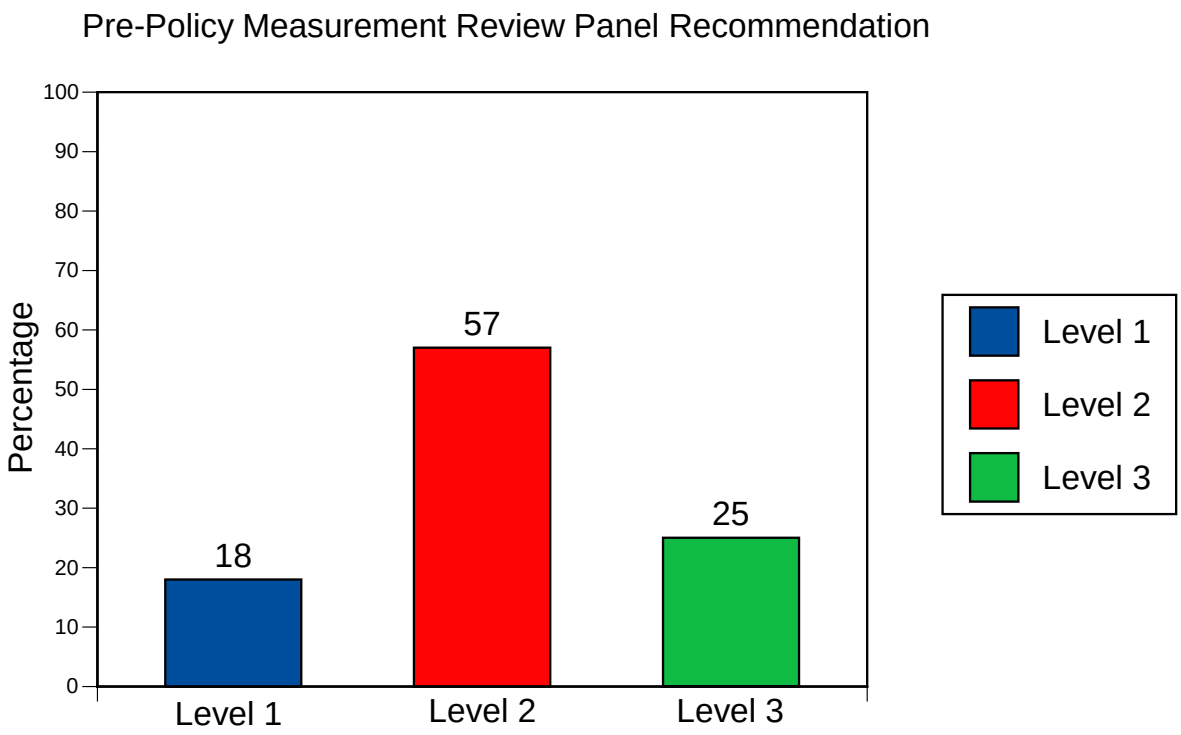

Figure 1. Mean Percentage of Students in Each Achievement Level Following Round Two

Evaluations

Exit surveys were completed by each panelist at the completion of the Pre-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 in Table 5, on the following page.

Table 5. Questionnaire Results for the Pre-Policy Measurement Review Panel Meeting (N=21)

Question	Mean	Median	Minimum	Maximum
1. The method for making recommendations on the ideal percent of students who should be classified in each achievement level was conceptually clear.		4.00	3	5
2. I had a good understanding of the design of the New York Regents Examination in Integrated Algebra.	4.25	4.00	3	5
3. I had a good understanding of the design for the other assessments presented such as NAEP.	3.70	4.00	2	5
4. After the first round of ratings, I felt comfortable with the method for making recommendations.	4.05	4.00	2	5
5. After the second round of ratings, I felt comfortable with the method for making recommendations.	4.50	5.00	2	5
6. I found the feedback on the recommendations of other panelists useful in making my own recommendations.	4.55	5.00	2	5
7. I found the feedback on the overall group recommendation useful in making my own recommendations.	4.50	5.00	3	5
8. I feel confident that the final cut score recommendations reflect the achievement levels associated with the New York State Regents Examination in Integrated Algebra.	4.30	4.00	2	5

Item Mapping Standard Setting

Two committees of New York educators were convened June 23 and June 24, 2008, in Albany to recommend standards for the Regents Examination in Integrated Algebra. The first, Committee A, had 31 educators, and the second, Committee B, had 30 educators. The item mapping procedure was applied to recommend the cut scores.

For Committee A, Dr. Paul Nichols of Pearson was the facilitator, and data analysis was provided by Dr. Robert Dolan of Pearson. For Committee B, Dr. Ye Tong of Pearson was the facilitator, and data analysis was provided by Morgan Hickey of Pearson. Replication of analyses was provided by Dr. Tom Hirsch from Assessment Evaluation Services (AES).

Panelists

All panelists provided voluntary demographic information. An example of the demographic questionnaire is provided in Appendix C.

A total of 31 panelists met in Committee A. A summary of the number of male and female panelists on Committee A is shown in Table 6, below.

Table 6. Numbers of Male and Female Panelists on Committee A

Male	Female
15	16

A summary of the ethnic representation of Committee A is provided in Table 7, below.

Table 7. Summary of the Ethnic Representation of the Panelists on Committee A

Caucasian	Hispanic	African-American	Other	Missing
20	2	4	4	1

Table 8, below, provides a summary of years of educational experience for the panelists on Committee A.

Table 8. Summary of Years of Educational Experience for the Panelists on Committee A

Mean	Median	Standard Deviation	Minimum	Maximum
15.40	14.00	9.73	4	41

A summary of district size for the panelists on Committee A is provided in Table 9, below.

Table 9. Summary of District Size for the Panelists on Committee A

Large	Medium	Small
15	5	10

A summary of district location for the panelists on Committee A is provided in Table 10, below.

Table 10. Summary of District Location for the Panelists on Committee A

Urban	Suburban	Rural
14	8	8

A total of 30 panelists met in Committee B. A summary of the number of male and female panelists on Committee B is shown in Table 11, below.

Table 11. Numbers of Male and Female Panelists on Committee B

Male	Female
20	10

A summary of the ethnic representation of the panelists on Committee B is provided in Table 12, below.

Table 12. Summary of the Ethnic Representation of the Panelists on Committee B

Caucasian	Hispanic	African-American	Other	Missing
21	2	3	2	2

Table 13, below, provides a summary of years of educational experience of the panelists on Committee B.

Table 13. Summary of Years of Educational Experience of the Panelists on Committee B

Mean	Median	Standard Deviation	Minimum	Maximum
15.53	14.00	8.77	2	33

A summary of district size for the panelists on Committee B is provided in Table 14, below.

Table 14. Summary of District Size for the Panelists on Committee B

Large	Medium	Small
14	11	5

Table 15, below, provides a summary of district location for the panelists on Committee B. Note that one panelist did not respond to this question.

Table 15. Summary of District Location for the Panelists in Committee B

Urban	Suburban	Rural
15	10	4

Method

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

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

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 will be 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 a quarter 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 Books

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. Selected-response items, such as true/false items and multiple-choice items, are presented only once in the book. Multiple-choice item pages will show the test item stem and alternatives, as well as the correct response. True/false item pages will show the test item and the correct response.

Constructed-response items are 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 page for each score point or item step will present the prompt and an example of student work awarded that particular score point. This example of student work should be a clear representation of performance at that score value. 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 items. Finally, the ordered item book would include 24 additional pages, 1 student work example for each of the 3 score points for each of the 8 items. The ordered item book would total 78 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. Under a Rasch model, empirical difficulty is simply the *b* parameter value for an item.

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 Integrated Algebra contains 30 selected-response items (multiple-choice items) and 9 constructed-response items. The selected-response items are weighted by 2 for scoring and the constructed-response items are weighted by 1. For the nine constructed-response items, three items have score range from 0 to 2, three items have score range from 0 to 3 and three items have score range from 0 to 4. The raw scores for the Regents Examination in Integrated Algebra range from 0 to 87.

Rasch Model

The Rasch model is used for all the Regents Examinations and the Partial Credit model is used for the constructed-response items for the assessment. Next, the Rasch model and the computation of empirical difficulty value with a response probability of 0.67 are discussed.

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 response probability is 0.50. To obtain the item parameter value and hence the corresponding θ value that will have a response probability of 0.67, modification needed to be conducted on the item parameters. Basically, the following equations need to be solved for b', the item difficulty, hence the ability level for a response probability of 0.67:

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

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

Solving this equation, we can have $b' = b + \ln 2 = b + 0.69315$. 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 items), 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 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 is more complicated than a simple addition factor, as is the case with dichotomously scored items. The idea was to produce the ability value that will yield a probability of 0.67 for a <u>given score category and above</u>. 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 response probability 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 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 also 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 H.

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. 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 from the test blueprint
- The key (unless the row corresponds to a score point for 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. For the augmented item map, the following additional information is commonly provided for each item:

- The percentage of the students who correctly answered the item (item *p*-value)
- The ability (in logits) required to answer the item correctly for a given response probability (RP) value

Item Mapping

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. This judgment is made within a given frame of reference, for a given response probability 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 typical RP values used with the bookmark method are 0.50 and 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 would answer correctly. For the immediately following item, panelists should judge that only 66 or fewer of 100 just Proficient examinees would respond correctly. For the immediately preceding item, panelists should judge 68 or more of 100 Just Proficient examinees would respond correctly. Panelists then mark that 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 a probability of 0.67 of answering the item correctly. The median of these Rasch values was then computed and was the cut score recommendation on the θ scale. To look up for the corresponding raw score cut, the raw to θ conversion table was used. The standard setting θ is 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 will be 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 on the Regents Examination in Integrated Algebra and items from the two anchor forms. The raw to θ conversion table was based on the operational Regents Examination in Integrated Algebra and had raw scores ranging from 0 to

87. 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 21. The item on page 21 had the θ value of 0.003 (see Appendix H, ordered item book). Next, we go to the raw to θ conversion table. A raw score of 40 corresponds to a θ value of -0.047; raw score of 41 corresponds to a θ value of 0.006. Per the NYSED's direction, the raw score cut then is 40. 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 Stem 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 Monday, June 23. The agenda for the standard setting conference is shown in Appendix D. The morning was devoted to introductions of the staff, to a description of standard setting, and to a description of the Regents Examination in Integrated Algebra.

Following the midmorning break, the two committees met in their separate conference rooms and began the process of reviewing achievement level descriptors. This process required several hours and resulted in a set of descriptors for each performance level (Level 1, Level 2, and Level 3).

Each committee then began the standard setting process. The item mapping procedure was the methodology used. In this procedure, panelists are asked to identify the item in an ordered item book that is the last item that a threshold student at a given level would be able to correctly answer. 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 2008 administration and anchor items from a field-test administration from 2007. Items were sorted from least to most difficult, using the Rasch item difficulty values.

Each ordered item book was accompanied by an item map containing the following:

1. Page number
2. Unique item identifier
3. Strand or content category
4. Correct option

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

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 five or six panelists. They were provided the cut scores 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, panelists were asked to think about the following:

- How similar are their 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?

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

In addition, panelists were provided a list of item *p*-values. Finally, panelists were provided a graphical display of the percentage of students in each achievement level on the basis of the mean cut scores from round one. For both

the *p*-values and the graphical display, a representative sample of approximately 144,000 students was used from the June 2008 operational administration of the Regents Examination in Integrated Algebra.

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 in 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 three tables combined, and noted the differences and similarities across tables, but reminded the panelists that a consensus was not required.

Next, panelists were provided a graphical display of the percentage of students in each achievement level on the basis of the mean cut scores from round two. A representative sample of approximately 144,000 students was used from the June 2008 operational administration of the Regents Examination in Integrated Algebra. Panelists were also provided with a graphical display of the percentage of students in each achievement level disaggregated for Grade 8 students and Grade 9 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 in each level.

Finally, panelists were asked to vote by a show of hands whether they favored immediately implementing the cut scores or possibly phasing them in over years.

Results

Table 16, below, summarizes the Committee A page number as well as raw score cuts for achievement Levels 2 and 3. These are the raw score recommendations from the committee.

Table 16. Committee A Cut Scores for Achievement Levels 2 and 3

	Page Number				Raw Score
	Mean	**Median**	**Minimum**	**Maximum**	
	Round 1				
Level 2	21	22	3	45	44
Level 3	64	64	39	81	63
	Round 2				
Level 2	21	21	9	43	40
Level 3	63	67	40	82	65
	Round 3				
Level 2	21	21	12	33	40
Level 3	65	67	45	79	65

Figure 2, below, shows the percentage of students in each performance level, using the Committee A cut scores after the round 3 final rating. The percentage of students in each achievement level is based on the operational sample of approximately 144,000 students collected in June 2008.

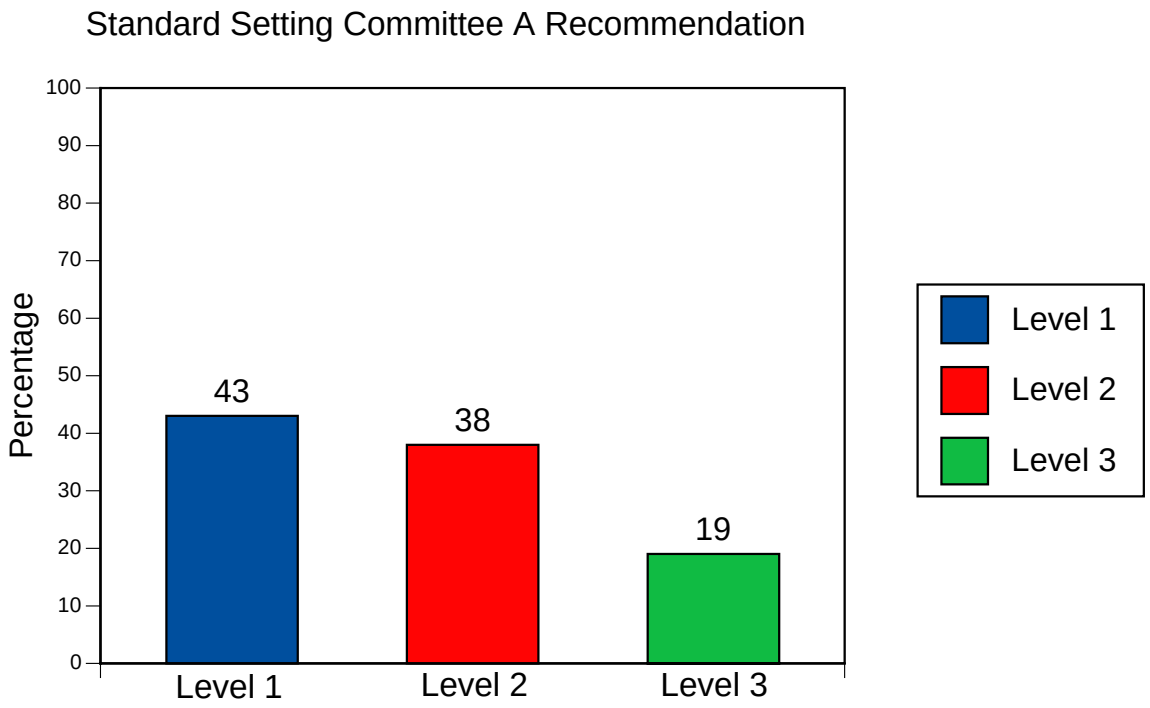

Figure 2. Percentage of Students in Each Achievement Level, Using Committee A Cut Scores After the Round Three Final Rating

Table 17, below, summarizes the Committee B page number as well as raw score cuts for achievement Levels 2 and 3 achievement. These are the raw score recommendations from the committee.

Table 17. Committee B Cut Scores for Achievement Levels 2 and 3

	Page Number				Raw Score
	Mean	**Median**	**Minimum**	**Maximum**	
	Round 1				
Level 2	36	40	15	53	52
Level 3	74	75	39	88	68
	Round 2				
Level 2	28	30	17	43	49
Level 3	71	69	51	83	65
	Round 3				
Level 2	25	27	17	40	48
Level 3	69	69	53	79	65

Figure 3, below, shows the percentage of students in each achievement level, using the Committee B cut scores after the round three final rating. The percentage of students in each achievement level is based on the operational sample of approximately 144,000 students collected in June 2008.

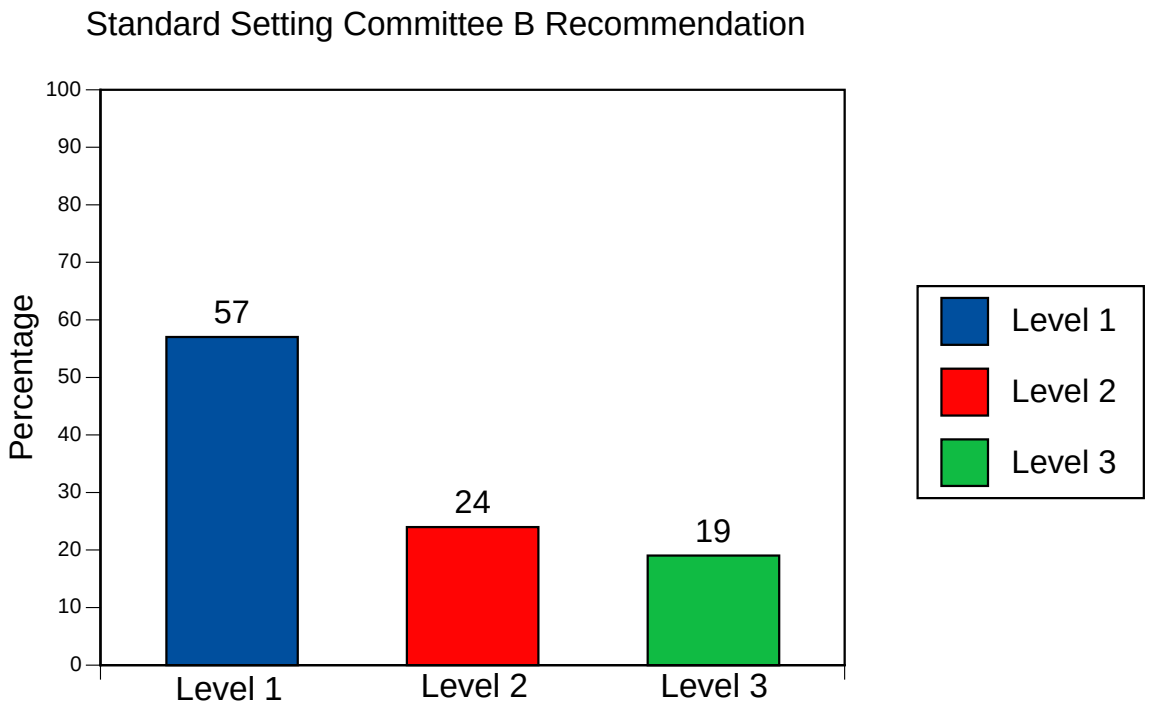

Figure 3. Percentage of Students in Each Achievement Level, Using Committee B Cut Scores After the Round Three Final Rating

Table 18, below, summarizes across both committees the page number as well as raw score cuts for achievement Levels 2 and 3. These are the raw score recommendations from the committee.

Table 18. Cut Scores Across Both Committees for Achievement Levels 2 and 3

	Page Number				Raw Score
	Mean	**Median**	**Minimum**	**Maximum**	
	Round 1				
Level 2	21	31	3	53	49
Level 3	64	71	39	88	66
	Round 2				
Level 2	22	24	9	43	46
Level 3	66	69	40	83	65
	Round 3				
Level 2	21	22	12	40	44
Level 3	66	69	45	79	65

Figure 4, below, shows the percentage of students in each achievement level, using the cut scores across both committees after the round three final rating. The percentage of students in each achievement level is based on the operational sample of approximately 144,000 students collected in June 2008.

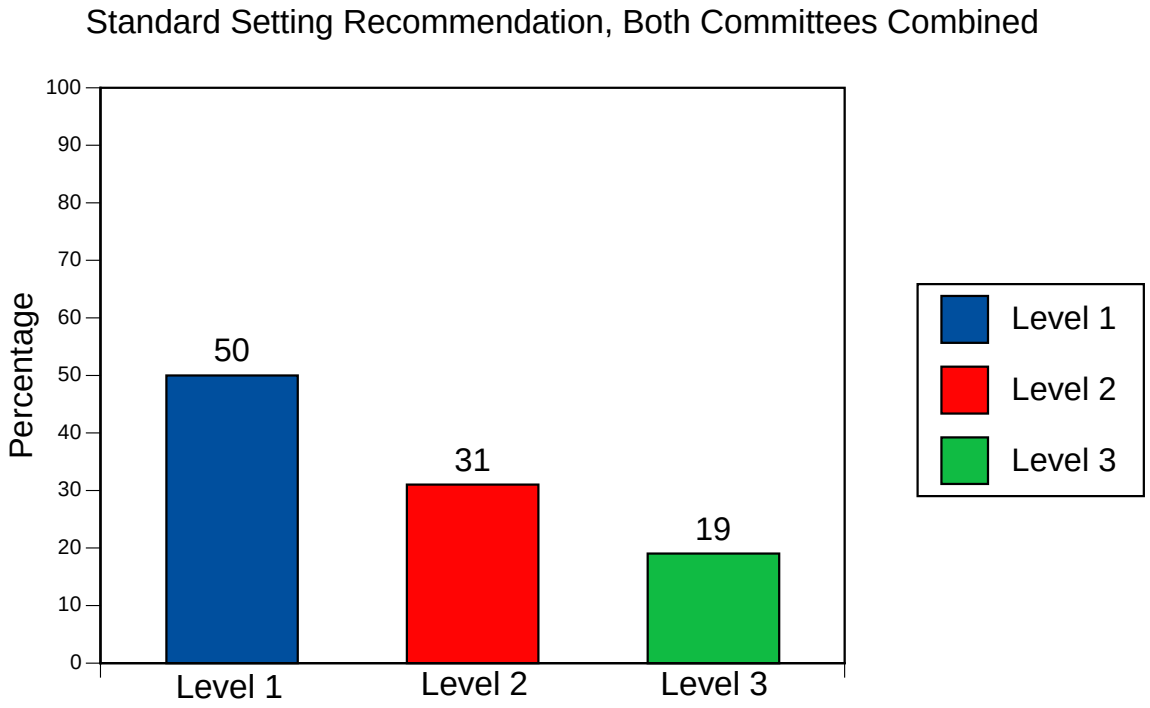

Figure 4. Percentage of Students in Each Achievement Level, Using Cut Scores Across Both Committees After the Round Three Final Rating

When the panelists were asked whether they favored immediately implementing the cut scores or possibly phasing them in over several years, two-thirds of the panelists voted for phasing them in over several years.

Panelist Variability

To describe the variability in panelists' judgments, a Generalizability Theory (G-Theory) study was performed. This information is used to determine how similar the cut scores might be if a different set of panelists or a 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 (three 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 results of the G-Theory study for the Committee A Level 2 cut score is shown in Table 19, below.

Table 19. The G-Theory Study for the Committee A Level 2 Cut Score

Variance Component	Estimated Variance Component	Applied Variance Component	Percent of Variance	Error Variance	Standard Error
Table	2.2525	2.2525	5		
Judge:Table	16.1692	16.1692	35		
Round	-0.7753	0.0000	0		
Table x Round	0.8782	0.8782	2		
Remaining	27.1749	27.1749	58		
				0.61868	0.78656

The results of the G-Theory study for the Committee A Level 3 cut score is shown in Table 20, below.

Table 20. G-Theory Study for the Committee A Level 3 Cut Score

Variance Component	Estimated Variance Component	Applied Variance Component	Percent of Variance	Error Variance	Standard Error
Table	13.4762	13.4762	44		
Judge:Table	5.8514	5.8514	19		
Round	-0.4282	0.0000	0		
Table x Round	1.9599	1.9599	6		
Remaining	9.5683	9.5683	31		
				2.75552	1.65998

The results of the G-Theory study for the Committee B Level 2 cut score is shown in Table 21, below.

Table 21. G-Theory Study for the Committee B Level 2 Cut Score

Variance Component	Estimated Variance Component	Applied Variance Component	Percent of Variance	Error Variance	Standard Error
Table	3.08463	3.08463	15		
Judge:Table	5.19778	5.19778	26		
Round	5.67500	5.67500	28		
Table x Round	2.21611	2.21611	11		
Remaining	4.13111	4.13111	20		
				0.66076	0.81287

The results of the G-Theory study for the Committee B Level 3 cut score is shown in Table 22, below.

Table 22. G-Theory Study for the Committee B Level 3 Cut Score

Variance Component	Estimated Variance Component	Applied Variance Component	Percent of Variance	Error Variance	Standard Error
Table	2.65324	2.65324	17		
Judge:Table	4.36667	4.36667	28		
Round	0.46528	0.46528	3		
Table x Round	-0.22750	0.00000	0		
Remaining	8.21778	8.21778	52		
				0.57802	0.76028

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 Committee A are shown in Table 23 and for Committee B in Table 24.

Table 23. Questionnaire Results for Committee A

Question	Mean	Median	Minimum	Maximum
1. The method for setting achievement levels, item mapping, was conceptually clear.	4.1	4.0	1	5
2. I had a good understanding of what the test was intended to measure.	4.2	4.0	3	5
3. I could clearly distinguish between student achievement levels.	3.9	4.0	3	5
4. After the first round of ratings, I felt comfortable with the achievement level setting procedure.	3.7	4.0	1	5
5. I found the feedback on item difficulty useful in setting achievement levels.	4.1	4.0	2	5
6. I found the feedback on the compared rating between judges useful in setting achievement levels.	4.3	4.5	2	5
7. I found the feedback on the percent of the students tested that would be classified at each performance level useful in setting achievement levels.	3.5	4.0	1	5
8. I feel confident that the final cut score recommendations reflect the achievement levels of the performance test associated with the New York State Regents Examination in Integrated Algebra.	2.8	3.0	1	5

Table 24. Questionnaire Results for Committee B

Question	Mean	Median	Minimum	Maximum
The method for setting achievement levels, item mapping, was conceptually clear.	4.3	5.0	1	5
I had a good understanding of what the test was intended to measure.	4.3	5.0	1	5
I could clearly distinguish between student achievement levels.	3.9	4.0	2	5
After the first round of ratings, I felt comfortable with the achievement level setting procedure.	3.6	4.0	2	5
I found the feedback on item difficulty useful in setting achievement levels.	3.8	4.0	2	5
I found the feedback on the compared rating between judges useful in setting achievement levels.	4.0	4.0	1	5
I found the feedback on the percent of the students tested that would be classified at each performance level useful in setting achievement levels.	3.9	4.0	2	5
I feel confident that the final cut score recommendations reflect the achievement levels of the performance test associated with the New York State Regents Examination in Integrated Algebra.	2.9	3.0	1	5

Post-Policy Measurement Review Panel

The Post-Policy Measurement Review Panel met on the afternoon of Tuesday, June 24, 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. The Post-Policy Measurement Review Panel was facilitated by Dr. Paul Nichols of Pearson, and Dr. Ye Tong of Pearson served as data analyst.

Panelists

This meeting was convened with 16 panelists from the Pre-Policy Measurement Review Panel.

Method and Procedure

The Post-Policy Measurement Review Panel meeting began approximately 30 minutes after the item mapping panelists had left. The agenda for the Post-Policy Measurement Review Panel is shown in Appendix F. The Post-Policy Measurement Review Panel meeting began with introductions of the facilitators and the NYSED staff. Panelists were then instructed that they were to review and integrate results from the Pre-Policy Measurement Review and the Item Mapping. The product of this activity would be final recommendations for the percentage of students in each achievement level that reflect the influence of both meetings.

Following these initial activities, panelists reviewed results from the Pre-Policy Measurement Review Panel. They then were given an explanation of the item mapping methodology. They then reviewed the results for both Committee A and Committee B from the item mapping meeting and 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. The recommendation form used by the Post-Policy Measurement Review Panel is shown in Appendix G.

Subsequently, 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 Pre-Policy Measurement Review Panel and Committee A and Committee B from the item mapping meeting. The panelists

were then asked, as a committee, to verbally make final adjustments to the percentage of students in each achievement level.

Finally, panelists were asked to vote with a show of hands whether they favored immediately implementing the cut scores or possibly phasing them in over several years.

Results

Table 25, below, summarizes, for the Post-Policy Measurement Review Panel, the panelists' recommendations for the ideal percentage of students who should be classified as Level 2 and above and the ideal percentage of students who should be classified as Level 3.

Table 25. Post-Policy Measurement Review Panel Results for the Mean Recommended Percentage of Students Classified as Level 2 and Above or Level 3

	Level 2 & Above	Level 3
Mean	74	21
Median	75	20
Maximum	90	30
Minimum	50	10

Figure 5, below, shows the percentage of students in each achievement level, using the mean recommendations from the Post-Policy Measurement Review Panel.

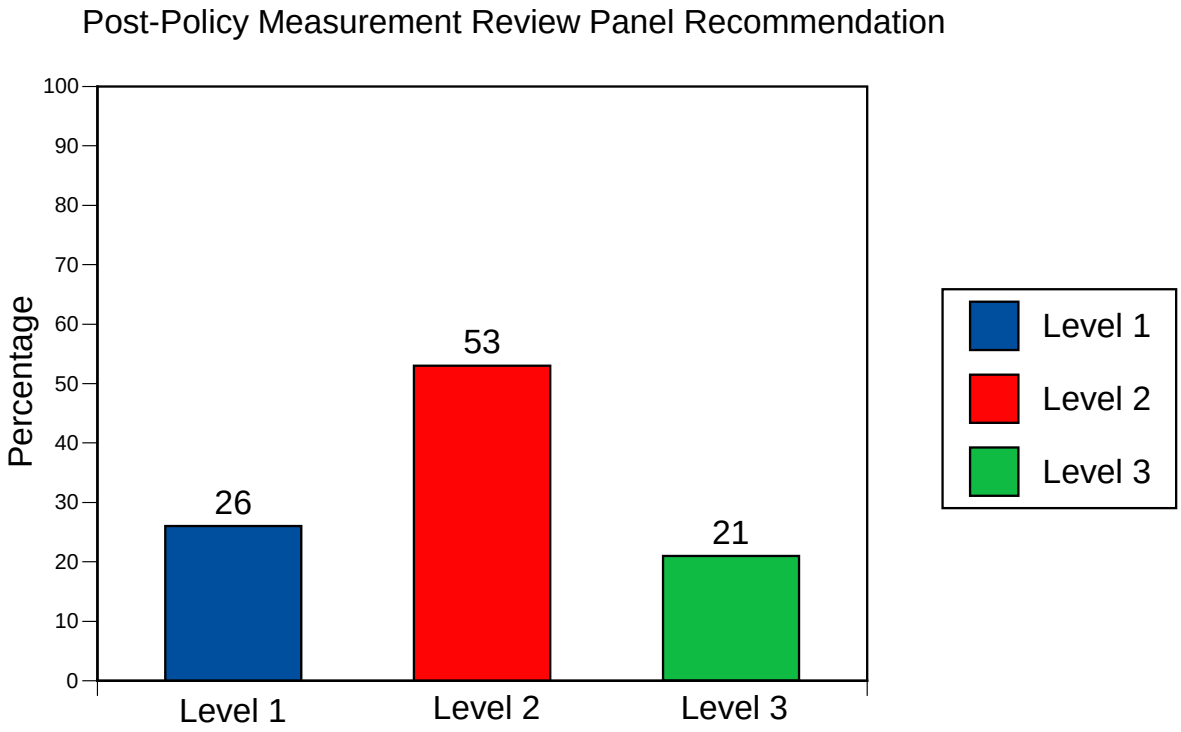

Figure 5. Mean Percentage of Students in Each Achievement Level, Using the Mean Recommendations from the Post-Policy Measurement Review Panel

When asked to verbally make final adjustments to the percentage of students in each achievement level, the majority of panelists indicated they favored decreasing the percentage of students at Level 1 by five to ten percentage points and increasing the percentage of students at Level 2 and above by five to ten percentage points.

When the panelists were asked whether they favored immediately implementing the cut scores or possibly phasing them in over several years, the panelists unanimously voted for phasing them in over several years. The panelists favored phasing in the raw score cut of 40 over a period of several years with the target percentage of students in each achievement level as shown in Figure 5.

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 26.

Table 26. Questionnaire Results for the Post-Policy Measurement Review Panel

Question	Mean	Median	Minimum	Maximum
1. The method for making recommendations on the ideal percent of students who should be classified in each achievement level was conceptually clear.	4.13	4.00	3	5
2. I had a good understanding of the results from the earlier meeting of the Measurement Pre-Policy review.	4.67	5.00	4	5
3. I had a good understanding of the results from the earlier Item Mapping Meeting.	4.47	5.00	3	5
4. After the first round of ratings, I felt comfortable with the method for making recommendations.	3.47	3.00	1	5
5. After the second round of ratings, I felt comfortable with the method for making recommendations.	3.36	4.00	1	5
6. I found the feedback on the recommendations of other panelists useful in making my second round recommendations.	3.93	4.00	2	5
7. I found the feedback on the overall group recommendations useful in making my second round recommendations.	4.07	4.00	3	5
8. I feel confident that the final cut score recommendations reflect the achievement levels associated with the New York State Regents Examination in Integrated Algebra.	3.33	3.00	1	5

Appendix A

Agenda for the Pre-Policy Review Panel Meeting

Agenda for Measurement Pre-Policy Review Meeting
New York State Regents Examination in Integrated Algebra
April 17, 2008

Registration	7:30-8:00
Opening Remarks Review of Agenda Security Forms Reimbursement	8:00-8:30
Welcome and Why You Are Here	8:30-9:00
Overview of Measurement Review Purpose Methodology	9:00-10:00
BREAK	10:00-10:15
Presentation of Related Assessments New York State Grade 6 Mathematics Test New York State Grade 7 Mathematics Test New York State Grade 8 Mathematics Test New York Regents Examination in Mathematics A New York Regents Examination in Mathematics B National Assessment of Education Progress (NAEP)	10:15-11:00
Break into Pre-assigned Groups	11:00
Discuss Assessment Data and Guiding Questions	11:00-11:45
Round 1 Recommendations Readiness Form Review Method Collect Recommendations	11:45-12:00
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 Readiness Form Review Method Collect Recommendations	2:30-2:45
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

Directions: Your task is to recommend the ideal percent of students who should be classified in each achievement level on the New York State Regents Examination in Integrated Algebra. First, recommend the ideal percent of students who should be classified as *Level 2 and above.* Next, recommend the ideal percent 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: _____

Table: _____

Round 1

Lower End of Range	
% Level 2 & Above	% Level 3
High End of Range	
% Level 2 & Above	% Level 3

Round 2

Lower End of Range	
% Level 2 & Above	% Level 3
High End of Range	
% Level 2 & Above	% Level 3

Appendix C

Demographic Questionnaire for the Item Mapping Committee Meetings

New York State Regents Examination in Integrated Algebra
Standard Setting
Panelist Information Sheet

Panelist ID Number: _____

Please provide the following demographic information that will be used to describe the general characteristics of the panelists who are recommending standards for the New York State Regents Examination in Integrated Algebra.

Your Current Position:

_____

_____

Courses / Grades Taught / Educational Experience:

_____

_____

_____

Gender (circle one): Male Female

Ethnicity:

_____

Years of Educational Experience:

_____

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

Small Medium Large

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

Urban Suburban Rural

Which best described the geographic location of your school 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

PEARSON

David Abrams
Assistant Commissioner
Office for Standards, Assessments and Reporting

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

Recommendations for Setting Achievement Levels for the Regents Examination in Integrated Algebra

Agenda

DAY 1 – June 23, 2008

Registration	7:30-8:00
Opening Remarks – Paul Nichols Introductions Why are you here? Agenda Confidentiality forms Reimbursement forms	8:00-8:45
Welcome and Overview – David Abrams Examination development Overview of standard setting process	8:45-9:30
Standard Setting Process – Paul Nichols Purpose Item Mapping Methodology	9:30-9:45
Overview of the Regents Examination in Integrated Algebra History Purposes Test Specifications	9:45-10:00
BREAK	10:00-10:15
Reconvene in committee rooms	10:15

Complete Integrated Algebra exam (individual committees)	10:15-11:15
Introduce Achievement Level Descriptors	11:15-11:30
Construct Achievement Level Descriptors Small Group Discussion	11:30-12:00
LUNCH (Train Table Leaders)	12:00-1:00
Construct Achievement Level Descriptors Small Group Discussion (continued) Large Group Discussion	1:00-2:15
BREAK	2:15-2:30
Assign panelist IDs	2:15-2:30
Overview of Standard Setting Item Mapping Ordered Item Booklet Item Map Ratings Forms	2:30-3:00
Practice Round	3:00-3:45
Round 1 Standard Setting Readiness Form Review Method Collect page number/item numbers	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 24, 2008

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	8:30-9:15
Round 2 Ratings Readiness Form Review Method Collect page number/item numbers	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 page number/item numbers	11:30-12:00
LUNCH	12:00-1:00
Feedback	1:00-1:15
Revisit Achievement Level Descriptors Review items near recommended cut scores 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

Appendix E

Recommendation Form for the Item Mapping Committee Meetings

Achievement Level Setting
New York State Regents Examination in Integrated Algebra
Item Position Recording Sheet

Panelist ID _____

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

Appendix F

Agenda for the Post-Policy Review Committee Meeting

University of the State of New York
State Education Department

PEARSON

David Abrams
Assistant Commissioner
Office for Standards, Assessments and Reporting

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

Measurement Post-Policy Review for the
New York State Regents Examination in Integrated Algebra
Agenda

AFTERNOON – June 24, 2008

Registration	2:30-3:00
Opening Remarks – Paul Nichols Introductions Agenda Confidentiality Forms Reimbursement Forms	3:00-3:10
Welcome and Overview – David Abrams Review of Standard Setting Process	3:10-3:20
Overview of Measurement Post-Policy Review Purpose Methodology	3:20-3:35
Review of the Measurement Pre-Policy Review	3:35-3:45
Review of Item Mapping	3:45-4:00
Comparison of Results	4:00-4:10

Round 1 Readiness Form Collect ratings form	4:10-4:25
BREAK	4:25-4:35
Presentation of Results	4:35-4:45
Round 2 Ratings Readiness Form Voting	4:45-4:55
Closing – David Abrams	4:55-5:00
End of Day Activities Check-in materials	5:00

Appendix G

Recommendation Form for the Post-Policy Review Committee Meeting

Rating Form
Post-Policy Review Workshop

Directions: Your task is to recommend the ideal percent of students who should be classified in each achievement level on the New York State Regents Examination in Integrated Algebra. First, recommend the ideal percent of students who should be classified as *Level 2 and above.* Next, recommend the ideal percent 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:

Round 1

Ideal percent of students	
% Level 2 & Above	% Level 3

Appendix H

Table H 1. Ordered Item Book

Page Number	Item ID	θ with RP 0.67	Page Number	Item ID	θ with RP 0.67
1	OPITEM_2	-1.457	46	ANCHOR2_9_3	0.745
2	ANCHOR2_2	-1.327	47	OPITEM_22	0.753
3	OPITEM_3	-1.317	48	ANCHOR2_7	0.773
4	OPITEM_1	-1.267	49	ANCHOR1_9_2	0.814
5	OPITEM_5	-1.137	50	ANCHOR2_10_1	0.832
6	OPITEM_4	-0.967	51	ANCHOR1_2	0.943
7	ANCHOR2_11_1	-0.828	52	ANCHOR2_8	0.943
8	OPITEM_7	-0.567	53	OPITEM_31_2	0.944
9	OPITEM_39_1	-0.530	54	OPITEM_34_2	0.949
10	OPITEM_6	-0.477	55	OPITEM_37_1	1.013
11	ANCHOR1_11_1	-0.250	56	ANCHOR1_7	1.013
12	ANCHOR1_8	-0.247	57	OPITEM_23	1.043
13	OPITEM_31_1	-0.211	58	ANCHOR1_11_3	1.081
14	ANCHOR1_11_2	-0.204	59	ANCHOR2_11_2	1.086
15	ANCHOR1_3	-0.097	60	ANCHOR2_9_4	1.086
16	ANCHOR1_9_1	-0.095	61	OPITEM_15	1.113
17	OPITEM_38_1	-0.092	62	OPITEM_21	1.133
18	OPITEM_36_1	-0.038	63	OPITEM_26	1.163
19	OPITEM_14	-0.037	64	ANCHOR1_11_4	1.215
20	OPITEM_10	-0.007	65	OPITEM_25	1.243
21	OPITEM_8	0.003	66	OPITEM_32_1	1.292
22	ANCHOR2_5	0.193	67	OPITEM_38_3	1.317
23	OPITEM_34_1	0.257	68	ANCHOR1_10_2	1.337
24	OPITEM_36_2	0.267	69	OPITEM_39_3	1.341
25	ANCHOR2_1	0.283	70	OPITEM_36_3	1.372
26	ANCHOR1_10_1	0.330	71	OPITEM_37_2	1.377
27	ANCHOR1_1	0.363	72	OPITEM_18	1.453
28	OPITEM_17	0.383	73	ANCHOR1_5	1.453
29	ANCHOR2_9_1	0.422	74	OPITEM_33_1	1.484
30	OPITEM_35_1	0.428	75	OPITEM_28	1.523
31	OPITEM_39_2	0.458	76	OPITEM_24	1.533
32	OPITEM_13	0.463	77	OPITEM_35_2	1.542
33	OPITEM_12	0.483	78	ANCHOR1_10_3	1.603
34	OPITEM_20	0.483	79	ANCHOR2_10_2	1.637
35	OPITEM_38_2	0.488	80	OPITEM_34_3	1.713
36	ANCHOR2_9_2	0.504	81	OPITEM_37_3	1.806
37	OPITEM_11	0.573	82	OPITEM_29	1.833
38	ANCHOR2_4	0.583	83	OPITEM_37_4	1.986
39	OPITEM_9	0.603	84	OPITEM_30	2.103
40	OPITEM_16	0.603	85	OPITEM_39_4	2.166
41	ANCHOR2_3	0.643	86	OPITEM_38_4	2.433
42	OPITEM_19	0.673	87	OPITEM_33_2	2.733
43	OPITEM_27	0.683	88	OPITEM_35_3	2.830
44	ANCHOR1_4	0.733	89	ANCHOR2_6	2.833
45	ANCHOR1_6	0.733	90	OPITEM_32_2	2.891

Table H 2. Raw to θ Conversion Table

Raw score	θ	CSEM	Raw score	θ	CSEM
0	-5.598	1.835	44	0.162	.227
1	-4.873	1.016	45	0.213	.227
2	-4.148	.730	46	0.265	.227
3	-3.710	.605	47	0.316	.227
4	-3.390	.531	48	0.368	.227
5	-3.134	.482	49	0.419	.227
6	-2.920	.446	50	0.471	.227
7	-2.734	.418	51	0.522	.227
8	-2.568	.396	52	0.574	.228
9	-2.419	.378	53	0.626	.228
10	-2.282	.362	54	0.679	.229
11	-2.156	.349	55	0.731	.230
12	-2.038	.338	56	0.784	.231
13	-1.927	.328	57	0.837	.231
14	-1.822	.319	58	0.891	.232
15	-1.723	.311	59	0.946	.234
16	-1.629	.304	60	1.000	.235
17	-1.538	.297	61	1.056	.237
18	-1.452	.291	62	1.112	.238
19	-1.368	.286	63	1.170	.240
20	-1.288	.281	64	1.228	.243
21	-1.211	.276	65	1.288	.245
22	-1.136	.272	66	1.348	.248
23	-1.063	.267	67	1.411	.252
24	-0.993	.264	68	1.475	.256
25	-0.924	.260	69	1.542	.260
26	-0.857	.257	70	1.611	.266
27	-0.792	.254	71	1.683	.272
28	-0.729	.251	72	1.759	.278
29	-0.667	.248	73	1.838	.286
30	-0.606	.246	74	1.923	.295
31	-0.546	.243	75	2.013	.306
32	-0.487	.241	76	2.110	.318
33	-0.430	.239	77	2.215	.332
34	-0.373	.237	78	2.330	.348
35	-0.317	.236	79	2.458	.367
36	-0.262	.234	80	2.602	.391
37	-0.207	.233	81	2.766	.420
38	-0.153	.232	82	2.958	.458
39	-0.100	.231	83	3.192	.510
40	-0.047	.230	84	3.489	.586
41	0.006	.229	85	3.905	.714
42	0.058	.228	86	4.608	1.005
43	0.110	.228	87	5.311	1.828