

**New York State Regents Examination in Physical
Setting/Earth Science
(Performance Test)
September 27-28, 2007**

Final Technical Report



**Prepared by:
Pearson
July 11, 2008**

TABLE OF CONTENTS

EXECUTIVE SUMMARY	3
Panelists	3
Method and Procedure	3
Results	4
Evaluations	6
FINAL REPORT	7
Panelists	7
Method and Procedure	8
Results	17
Evaluations	19
REFERENCES	20
TABLES	21
FIGURES	24
Appendix A	27
Appendix B	29
Appendix C	34
Appendix D	38
Appendix E	41
Appendix F	43
Appendix G	45
Appendix H	46

**New York State Regents Examination in Physical Setting/Earth Science
Performance Test
September 27-28, 2007**

EXECUTIVE SUMMARY

A committee of 26 New York State educators met on September 27 and September 28, 2007, in Malta, New York, to recommend cut scores for the New York State Regents Examination in Physical Setting/Earth Science (Performance Test). This executive summary will provide a brief description of the procedure and results of the standard setting study. An executive summary is delivered within several days of the conclusion of the study. A more detailed description of the study, including an explanation of the method and a round-by-round report of the results, is presented in the technical report.

Panelists

A total of 26 panelists participated in the standard setting for the New York State Regents Examination in Physical Setting/Earth Science (Performance Test).

A summary of panelist gender and ethnicity is provided in Table 1. A more detailed summary of panelist demographic information is provided in the technical report.

Table 1. A summary of gender and ethnicity data for the committee members.

Gender		Ethnicity			
Male	Female	Caucasian	Hispanic	African-American	Asian-American
9	17	19	3	1	3

Methodology

Panelists used an item mapping methodology, sometimes referred to as a Bookmark approach (Lewis, Mitzel & Green 1996; Mitzel, Lewis, Patz, & Green, 2001), to recommend cut scores for the New York State Regents Examination in Physical Setting/Earth Science (Performance Test). Cut scores are points on the score scale that define the boundaries of achievement levels.

Procedure

The standard setting conference began on Thursday, September 27. The morning was devoted to introductions of the staff, to a description of standard setting, and to a description of the New York State Regents Examination in Physical Setting/Earth

Science (Performance Test). As part of this process, panelists attempted the New York State Regents Examination in Physical Setting/Earth Science (Performance Test).

Following the midmorning break, the committees began the process of creating achievement level descriptors. This process required several hours. The result from creating achievement level descriptors was a set of descriptors for each performance level (0-64, 65-84 and 85-100).

During the afternoon, the committee began the standard setting process. During Round 1, panelists were asked to review items in the ordered item book and make a judgment as to the likelihood of threshold examinees answering an item correctly. Panelists were instructed to identify the last item in an ordered item book that a threshold student at a given achievement level would have a response probability of at least 0.67 of answering correctly. Panelists began with achievement level 65-84 and then moved to achievement level 85-100. Panelists recorded their judgments on a ratings sheet.

Following Round 1, panelists met in small groups of 5 or 6 panelists. They were provided the cut scores for each panelist based on the Round 1 ratings in addition to the mean, median, minimum and maximum cut score at each level for that table. During Round 2, panelists were again asked to review items in the ordered item book and make a judgment as to the likelihood of threshold examinees answering an item correctly.

Following Round 2, panelists received three kinds of feedback. First, panelists received the same feedback for each table that was provided following Round 1. Second, panelists were given the mean, median, minimum and maximum cut scores for the committee (across tables). Third, panelists were provided a graphic display of the percent of students in each achievement level from the June, 2006, Earth Science Test, based on the entire testing population at that administration. Panelists were also provided with a graphic display of the percent of students in each achievement level on the written section of the science test based on a representative sample of students from June, 2006.

Finally, panelists were provided a graphic display of the impact of using the median cut score for all students. The impact data graphic representation provided panelists with information on what percentages of students are at each performance level for the field test sample of 630 students on the science performance test. This data was collected in the summer of 2006.

For Round 3, panelists were again asked to review the placement of their bookmarks in the ordered item book and make a judgment as to the likelihood of threshold examinees answering an item correctly. They recorded their judgment on their ratings sheet.

Results

Cut scores at each round for achievement levels II and III are provided below in Tables 2 and 3, respectively.

Table 2. Cut scores by round for the Level II (65-84) achievement level.

Round	Item Number				Theta	Raw
	Mean	Median	Minimum	Maximum		
1	16	15	9	32	0.77815	12
2	14	9	5	23	0.38415	10
3	12	9	5	23	0.38415	10

Table 3. Cut scores by round for the Level III (85-100) achievement level.

Round	Item Number				Theta	Raw
	Mean	Median	Minimum	Maximum		
1	33	33	16	39	1.50415	15
2	31	33	20	36	1.50415	15
3	30	33	18	35	1.50415	15

Figure 1 shows the percentage of students in each performance level using the cut scores after the Round 3 final rating. The percentage of students in each performance level is based on the field test sample of 630 students on the Physical Setting/Earth Science Performance Test. This data was collected in the summer of 2006.

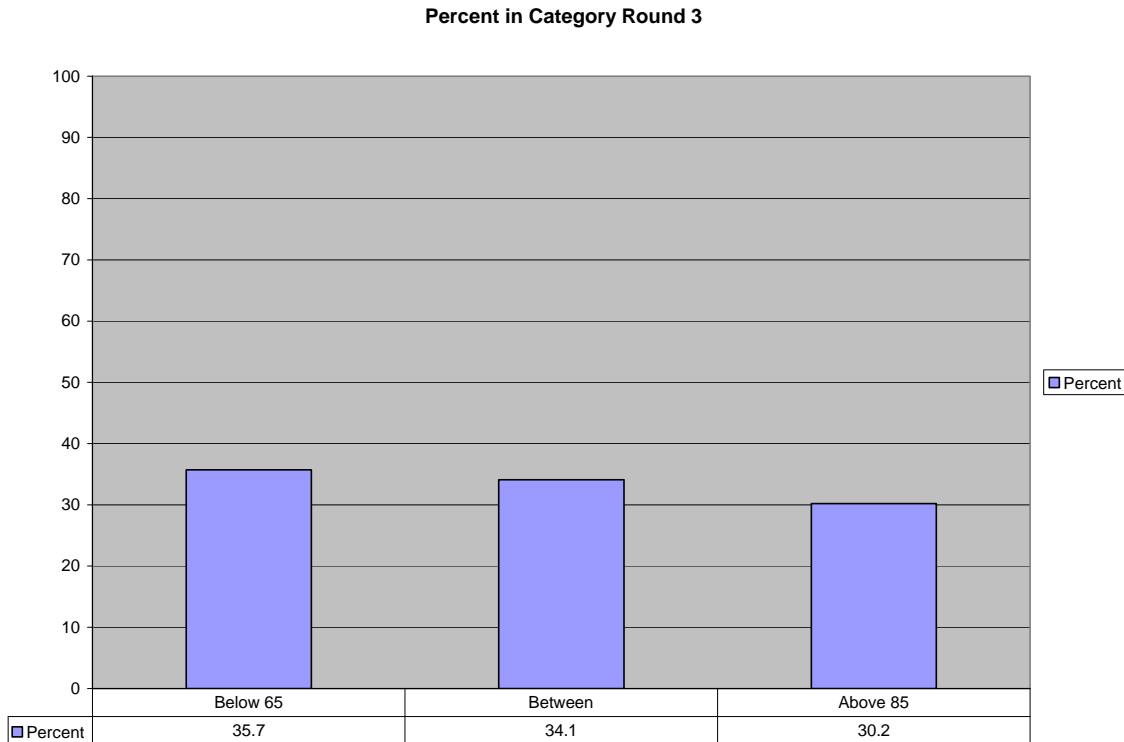


Figure 1. Percent of students at each achievement level when applying Round 3 cut scores.

Evaluations

Evaluations were administered to panelists following completion of the study. They responded to each question using a scale from 1 (**totally disagree**) to 5 (**totally agree**). See Table 3 below for the results of this survey.

Table 3. Exit survey results

Question	Mean	Median	Minimum	Maximum
The method for setting achievement levels, item mapping, was conceptually clear.	4	4	3	5
I had a good understanding of what the test was intended to measure.	5	5	3	5
I could clearly distinguish between student achievement levels.	4	4	2	5
After the <u>first</u> round of ratings, I felt comfortable with the achievement level setting procedure.	4	4	2	5
I found the feedback on item difficulty useful in setting achievement levels.	4	4	2	5
I found the feedback on the compared rating between judges useful in setting achievement levels.	4	5	2	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.	4	4	1	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 Physical Setting/Earth Science.	4	4	1	5

FINAL REPORT

A panel of 26 New York educators met on September 27 and 28, 2007, in Malta, New York, to set standards for the New York State Regents Examination in Physical Setting/Earth Science (Performance Test). Pearson representatives Dr. Paul Nichols and Dr. Ye Tong, conducted the standard setting study and used an item mapping procedure to recommend criteria that determine student achievement levels. A detailed description of the study and the criteria recommendations are provided in this report.

Panelists

An important component of a valid, legally defensible standard setting plan is to obtain the very best judgments from people in the best position to make the judgments. Consequently, standard setting panelists should be subject matter experts, understand the examinee population, be familiar with the curriculum, have knowledge of the instructional environment, have an appreciation of the consequences of the standards, and be representative of all stakeholder groups. This is a unique clustering of knowledge and skills and it may be difficult to gather a panel where every member meets every criterion.

The standard setting panel should be composed of stakeholders in the assessment process including teachers, curriculum specialists, school administrators, parents and community members. All standard setting panelists should have direct experience with students at the grade level for which standards are being set or with curriculum materials relevant for that grade level. This broad representation on the standard setting panel provides that the standards set for proficiency will have careful scrutiny from a broad range of constituents of education.

The panelists invited to the standard setting meeting for the New York State Regents Examination in Physical Setting/Earth Science (Performance Test) had to meet certain criteria. Panelists invited were certified Earth Science teachers who had been teaching at least three years. Panelists were recruited from all over the state so that the panel would represent veteran and newer teachers and different ethnic backgrounds, ELS and special needs. The attempt was made to recruit panelists from the Big Five cities and rural and suburban schools through science coordinators or science chairpersons.

A summary of panelist demographic information is provided in Tables 1 through 4, immediately following the Final Report. All panelists provided voluntary demographic information, using the form shown in Appendix A.

Methodology

Panelists used an item mapping methodology, sometimes referred to as a Bookmark approach, to recommend standards of the New York State Regents Examination in Physical Setting/Earth Science (Performance Test) (Lewis, Mitzel & Green 1996; Mitzel, Lewis, Patz, & Green, 2001).

The item mapping methodology is typically conducted using the following materials:

- Achievement level descriptors (ALDs);
- Ordered item books; and,
- Item maps.

A description of each of these is provided to give a background for a description of the item mapping methodology. The description of each of these components is taken from Nichols, O'Malley, Twing and Mueller (in preparation). 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 goals. 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, called achievement level descriptors (ALDs).

Generally, achievement levels represent a broad range of achievement (Mitzel, et al, 2001). For example, more than a quarter of the students in a grade level for a state may be classified as failing with 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 performance 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 barely Proficient or barely Advanced students know and can do are created. These students that are barely Proficient or barely 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 barely Proficient or barely Advanced students know and can do play a central role in standard setting. The panelists are instructed to use these ALDs of what barely Proficient or barely Advanced students know and can do as the frame of

reference for each judgment. The construct being measured is the panelists' representation of barely Proficient or barely 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 know 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 Bookmark 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 to the most difficult items.

The ordered item book includes 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 four-point rubric (0 to 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, one page for each of the 30 multiple choice items. In addition, the ordered item book would include 24 pages, one page for each of the three score points for the eight items. Finally, the ordered item book would include 24 additional pages, one student work example for each of the three score points for the eight items. The ordered item book would total 78 pages.

Sometimes an ordered item book is constructed 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 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 for should be greater for higher score points than for lower score points. A score point of at least three will be more difficult to obtain than a score point of at least two.

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 that 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; and,
- The key (unless the row corresponds to a score point for a constructed response item).

Following the first round 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 percent of the students who correctly answered the item (item p value); and,

- The ability (in logits) required to answer the item correctly for a given 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 (just Basic, just Proficient and just Advanced) 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 a given score point. The likelihood is commonly referred to as a response probability (RP) value. The typical RP values used with the bookmark method are 0.50 and 0.67. Panelists may be instructed to conceive 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 out 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 percent 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 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 least to most difficult items. Keeping in mind the ALD, panelists are instructed to identify the last item or score point that 67 out of 100 threshold students would answer correctly. For the immediately following item, panelists should judge that only 66 or fewer out of 100 Just Proficient examinees would respond correctly. For the immediately preceding item, panelists should judge 68 or more out of 100 Just Proficient examinees would respond correctly. Panelists then mark that page in the ordered item book, often using a sticky note, and record the item identifier on a record sheet.

Methodological strengths

The item mapping method has several features that make the method an appealing standard setting approach. First, the item mapping method can be used with a mixed-format assessment (Cizek & Bunch, 2007). 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 panelists complete within the item mapping method may be relatively less challenging than the panelists' task under other standard setting methods (Mitzel et al., 2001). Proponents of the item mapping method argue that panelists are required to make relatively few judgments compared to 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 a judgment for each item.

In addition, panelists using the item mapping method are required to spend relatively less time reviewing the test items (Cizek & Bunch, 2007). 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.

However, the item mapping method does require that items have been previously scaled using Item Response Theory. Before an item mapping procedure can be conducted, substantial work must be done including collecting student responses and calibrating and scaling items. Student responses may be collected through either a field test or operational administration. An operational administration is likely to provide a larger number of responses collected under more realistic conditions than a field test.

Procedure

In this section, we document the operationalization of the item mapping methodology as used to recommend cut scores for the New York State Regents Examination in Physical Setting/Earth Science (Performance Test).

The standard setting conference began on Thursday, September 27. The agenda for the meeting is shown in Appendix B. The morning was devoted to introductions of the staff, to a description of standard setting, and to a description of the New York State Regents Examination in Physical Setting/Earth Science (Performance Test).

As part of this process, panelists attempted the New York State Regents Examination in Physical Setting/Earth Science (Performance Test). The examination stations were set up and panelists completed the items associated with each station. Panelists worked in small groups and the activities were timed. Panelists were instructed that the purpose of taking the test was to familiarize themselves with the cognitive demands of the items. Panelists were instructed to take notes on sources of item difficulty.

Following the midmorning break, the committees began the process of creating achievement level descriptors. This process required several hours. The result from creating achievement level descriptors was a set of descriptors for each threshold performance level (65-84 and 85-100).

During the afternoon, the committee began the standard setting process. The item mapping procedure was the judgmental process 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 an RP of at least 0.67 of answering correctly (Huynh, 2006).

Ordered item books contained the performance test items as well as non-performance Earth Science items. These items were ordered from least difficult to most difficult according to Rasch item difficulty values. These values were calculated using data collected from NYS Earth Science students during the May, 2006 field test administration.

Each ordered item book was accompanied by an item map containing the ordered item book page number, unique item identifier, strand or content category, respective p-value, and correct option.

During Round 1, panelists were asked to review items in the ordered item book and make a judgment as to the likelihood of threshold examinees answering an item correctly. Panelists were instructed to identify the last item in an ordered item book that a threshold student at a given achievement level would have a response probability of at least 0.67 of answering correctly. Panelists began with achievement level 65-84 and then moved to achievement level 85-100. Panelists recorded their judgments on a ratings sheet.

Following Round 1, panelists met in small groups of 5 or 6 panelists. They were provided the cut scores for each panelist based on the Round 1 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 are 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 conceptualization of these threshold students?

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

During Round 2, panelists were again asked to review items in the ordered item book and make a judgment as to the likelihood of threshold examinees answering an item correctly.

Following Round 2, panelists received three kinds of feedback. First, panelists received the same feedback for each table that was provided following Round 1. Second, panelists were given the mean, median, minimum and maximum cut scores for the committee (across tables). The facilitator led the discussion with all tables combined. The facilitator noted the differences and similarities across tables but reminded the panelists that consensus was not required.

Third, panelists were provided a graphic display of the percent of students in each achievement level from the June, 2006, Earth Science Test, based on the entire testing population at that administration. Panelists were also provide with a graphic display of

the percent of students in each achievement level on the written section of the Earth Science test based on a representative sample of students from June, 2006.

Finally, panelists were provided a graphic display of the impact of using the median cut score for all students. The impact data graphic representation provided panelists with information on what percentages of students are at each performance level for the field test sample of 630 students on the Physical Setting/Earth Science Performance Test. This data was collected in May, 2006. Panelists were given time to discuss, within the big group, the appropriateness of the committee level cut scores given the proportion of students that would fall in each level.

For Round 3, panelists were again asked to review the placement of their bookmarks in the ordered item book and make a judgment as to the likelihood of threshold examinees answering an item correctly. They recorded their judgment on their ratings sheet.

After the Round 3 rating sheets were collected, Pearson staff members analyzed data and produced the final cut score recommendations. The panelists reconvened and were presented the final cut score recommendations. After reviewing this information panelists were asked to revisit the performance level descriptors created earlier in the process and complete a content analysis of the items in the ordered item booklet surrounding the recommended cut scores. Panelists were asked to review five items above each cut score and five items below each cut score, and determine the knowledge and skills necessary to successfully complete these items. They were then asked to review the performance level descriptors and make any modifications necessary to ensure that the knowledge and skills identified in the items surrounding the cut scores are represented in the descriptors. Panelists began the process by working and discussing in small groups of 5 to 6 people. Each group then presented their ideas to the larger group. Finally, the larger group worked together to make edits to and finalize the performance level descriptors.

The panelists were then asked to complete a short questionnaire, evaluating the standard setting process. The questionnaire asked about panelists' level of comfort with the standard setting procedure, their understanding of the performance levels and their satisfaction with final cut scores (see Appendix E).

Ordered item book

The ordered item books were constructed from the field test form available from May, 2006 field test. In addition, a set of linking items were included from the written portion of the test. Items were ordered by Rasch item difficulty values calculated from the field test data for the performance portion of the items. Items were then sorted from least to most difficult.

The Rasch Item Response Theory (IRT) model was used for psychometric analysis for the New York State Regents exams. When it is a dichotomous item, the Rasch model can be defined the following:

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

From the field test data, by definition, the item difficulty parameter b was calculated for a student with the same ability as the b parameter to have a probability of 0.50 to answer the item correctly. To obtain the item parameter value hence the corresponding θ value that will have a response probability of 0.67, modification needed to be conducted on the field test 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 field test item parameters (dichotomous items only) for the items to be included in the ordered item book.

When it is a polytomously scored item, the formulas are a bit more complicated. The IRT partial credit model (PCM) is used to analyze polytomously scored constructed response items for the New York State Regents exams. 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 $x = 0, 1, \dots, m_i$. D_{ij} values were available from field test analysis 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. Beretvas (2004) provides some detailed procedures to use for identifying the location of the item difficulty under various RP values. Again, modification was applied to the two-point constructed response items to obtain their difficulty values under RP 0.67, and the ordered item book was constructed.

Each ordered item book was accompanied by an item map. The item map for the New York State Regents Examination in Physical Setting/Earth Science (Performance Test) is shown in Appendix C. The item map contains eight pieces of information:

1. Page number
2. A unique item identifier
3. Strand or content category
4. P value
5. Correct option

Cut Score Computation

The cut score at each achievement level was determined by computing the median item number across panelists at a given grade level and its associated Rasch item difficulty with RP value of 0.67. This represents the minimum raw score that an examinee must attain to be classified at the particular level. When “translated” from item number in the ordered item book to the Rasch value and hence the raw score, cuts are usually computed to be between two raw scores. In the final report, all cut scores are rounded down to the raw score point closest to the theta value without going over, to give students the benefit of the doubt.

As mentioned before, the ordered item book contained 42 pages representing 42 score points. The ordered item book (OIB) was constructed from the field test form available from the May, 2006 field test and a set of anchor items from the written portion of the test. The item bookmark difficulty location was determined for a RP value of 0.67 using the procedures described by Beretvas (2004).

Panelists were instructed to begin at the front of the OIB and work page-by-page through the book. Panelists reported, using a rating sheet, in the last item that was judged a student who was “just Level II” or “just Level III” would have at least a 0.67 chance of answering correctly.

The median panelist rating was computed for each achievement level. Using that median ability value, the corresponding raw score was identified on the 22 point Earth Science (Performance Test). Based on earlier discussions with NYSED, the raw score was identified that was nearest to, but less than, the median ability value. The ability value associated with that raw score was also identified.

For example, at round 3, the median page number for 65-84 achievement level from the panelists’ rating was 9. The item on page 9 was item 11 on the performance test. From field testing analysis, this item had a difficulty value of -0.309. The ability value corresponding to a response probability for this item was computed to be 0.38415 (it was a result of $-0.309 + 0.69315$, see description before this section). Therefore, the theta cut for 65-84 achievement level, based on round 3 results, was 0.38415. A raw-score-to-theta distribution was also produced for the Performance Part D test, using the item parameters obtained from the field testing. This raw-score-to-theta distribution has raw scores ranging from 0 to 22 and it is provided in Appendix H. As can be observed from this table, a raw score of 10 corresponded to a theta value of 0.288 and a raw score of 11 corresponded to a theta value of 0.496. The theta cut for 65-84 was 0.38415. To follow the direction from the State and to give students the benefit of the doubt, the raw score of 10 was assigned to be the raw score cut for the performance level 65-84 based on round 3 results. Other performance levels and other rounds raw cut scores are similarly identified.

After the Round 3 rating sheets were collected, Pearson staff members analyzed the data and produced the final cut score recommendations. The panelists reconvened and were presented the final cut score recommendations.

Results

The results are summarized for each round. The cut scores for each level are presented as well as the impact data. Raw scores are the number of points earned by a student. Theta values are scale values, reported in logits, that result from the application of the Rasch item response theory model.

Round 1

Table 5 summarizes the item number and raw score for the Level II (65-84) achievement level and Level III (85-100) achievement level cut scores from Round 1. These are panelist ratings aggregated across the committee.

Figure 1 shows the percentage of students in each performance level using the cut scores after the Round 1 ratings. This data was never shown to the panelists. The percentage of students in each performance level is based on the field test sample of 630 students on the Physical Setting/Earth Science Performance Test. This data was collected in the May, 2006.

Round 2

Table 6 summarizes the item number and raw score for the Level II (65-84) achievement level and Level III (85-100) achievement level cut scores from Round 2. Again, the values reported in the table are panelist ratings aggregated across the committee.

Figure 2 shows the percentage of students in each performance level using the cut scores after the Round 2 ratings. This data was shown to the panelists as impact data. The percentage of students in each performance level is based on the field test sample of 630 students on the science performance test. This data was collected in May, 2006.

Round 3

Table 7 summarizes the item number as well as raw score cuts for Level II (65-84) and Level III (85-100) achievement levels following Round 3. These are the raw score recommendations from the committee.

Figure 3 shows the percentage of students in each performance level using the cut scores after the Round 3 final rating. The percentage of students in each performance level is

based on the field test sample of 630 students on the Earth Science (Performance Test). This data was collected in May, 2006.

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 were 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 urGENOVA program (Brennan, 2001). 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. Variances components for tables (σ_{Tables}^2) and judges within tables ($\sigma_{Judges:Tables}^2$) were computed. Computation of the standard errors were made using the following formula (Lee & Lewis, 2001):

$$SE_{cut} = \sqrt{\frac{\sigma_{Tables}^2}{N_{Tables}} + \frac{\sigma_{Judge:Table}^2}{N_{Judges} \bullet N_{Tables}} + \frac{\sigma_{Error}^2}{3N_{Tables} \bullet N_{Judges}}}$$

Because round was treated as a fixed facet, its variance component was not included in the error term. σ_{error}^2 was a confounding term and included the variance from the interaction between tables and judges within tables as well as variances unexplained by the defined facets. The sample size in the equation referred to the sample size likely to occur in the Decision Study (D-Study). Without loss of generality, the sample sizes for the D-Study were assumed the same as the sample size in the G-Study. Standard errors were computed for each of the two recommended cut scores. Different patterns of variance component estimates and, hence, standard errors for cut scores were anticipated for different cut scores (Lee & Lewis, 2001).

Table 8 provides the standard errors for the cut scores from round 3. Error sources were the variability of judges and tables. Round was treated as a fixed factor.

The standard errors for Level II and Level III were applied to the cut scores from Round 3. The intervals of plus or minus one standard error around the cut score for Level II and Level III are shown in Table 9.

Evaluations

Evaluations were administered following the completion of standard setting. An evaluation was completed by each panelist. Panelists answered each question using a scale from 1 to 5, 1 being “**totally disagree**” and 5 being “**totally agree**”. The survey questions and the results are shown in Table 10.

REFERENCES

- Beretvas, S. N. (2004). Comparison of Bookmark Difficulty Locations Under Different Item Response Models. Applied Psychological Measurement, 28, 25-47.
- Brennan, R. L. (2001). Manual for urGENOVA. Iowa City, IA: Iowa Testing Programs, University of Iowa.
- Huynh, H. (2006). A clarification on the response probability criterion RP67 for standard settings based on bookmark and item mapping. Educational Measurement: Issues and Practice, 25 (2), 19-20.
- Lee, G, & Lewis, D. (2001, April). A Generalizability Theory approach toward estimating standard errors of cut scores set using the Bookmark standard setting procedure. Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.
- .

Tables

Table 1. A summary of gender and ethnicity data for the committee members.

Gender		Ethnicity			
Male	Female	Caucasian	Hispanic	African-American	Asian-American
9	17	19	3	1	3

Table 2. A summary of panelists' responses to district size.

District Size		
Large	Medium	Small
12	10	4

Table 3. A summary of panelists' responses to district location.

District Location		
Urban	Suburban	Rural
16	7	3

Table 4. A summary of panelists' responses to district geographic location.

Geographic Location				
North	South	East	West	Central
1	12	3	3	7

Table 5. Cut scores for Round 1.

Achievement Level	Item Number				Theta	Raw Score
	Mean	Median	Minimum	Maximum		
II	16	15	9	32	0.77815	12
III	33	33	16	39	1.50415	15

Table 6. Cut scores for Round 2.

Achievement Level	Item Number				Theta	Raw Score
	Mean	Median	Minimum	Maximum		
II	14	9	5	23	0.38415	10
III	31	33	20	36	1.50415	15

Table 7. Cut scores for Round 3.

Achievement Level	Item Number				Theta	Raw Score
	Mean	Median	Minimum	Maximum		
II	12	9	5	23	0.38415	10
III	30	33	18	35	1.50415	15

Table 8. The standard errors for the cut scores from Round 3.

Level	
Level II (65-84)	Level III (85-100)
0.5356	0.3655

Table 9. Intervals of plus or minus one standard error around the cut score.

Achievement Level	Raw Score	Standard Error	-1 Standard Error	+1 Standard Error
II	10	0.5356	9.4644	10.5356
III	15	0.3655	14.6345	15.3655

Table 10. The questionnaire results for the standard setting committee.

Question	Mean	Median	Minimum	Maximum
The method for setting achievement levels, item mapping, was conceptually clear.	4	4	3	5
I had a good understanding of what the test was intended to measure.	5	5	3	5
I could clearly distinguish between student achievement levels.	4	4	2	5
After the <u>first</u> round of ratings, I felt comfortable with the achievement level setting procedure.	4	4	2	5
I found the feedback on item difficulty useful in setting achievement levels.	4	4	2	5
I found the feedback on the compared rating between judges useful in setting achievement levels.	4	5	2	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.	4	4	1	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 Physical Setting/Earth Science.	4	4	1	5

Figures

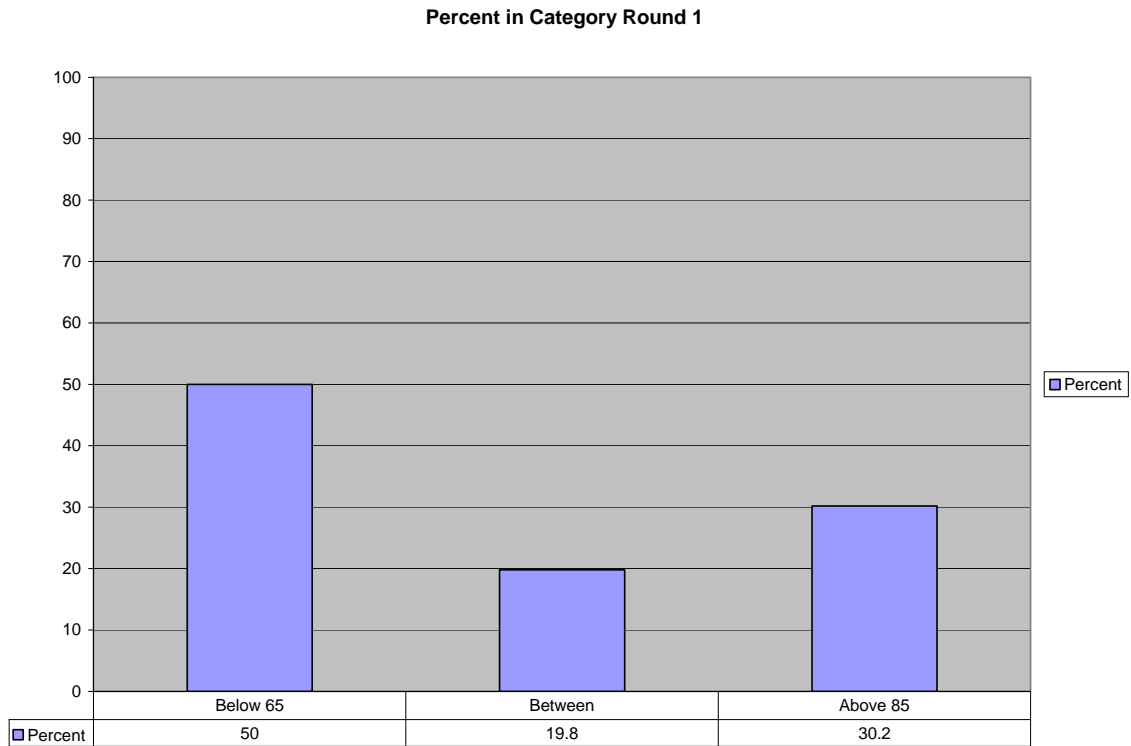


Figure 1. The percentage of students in each performance level using the cut scores after Round 1.

Percent in Category Round 2

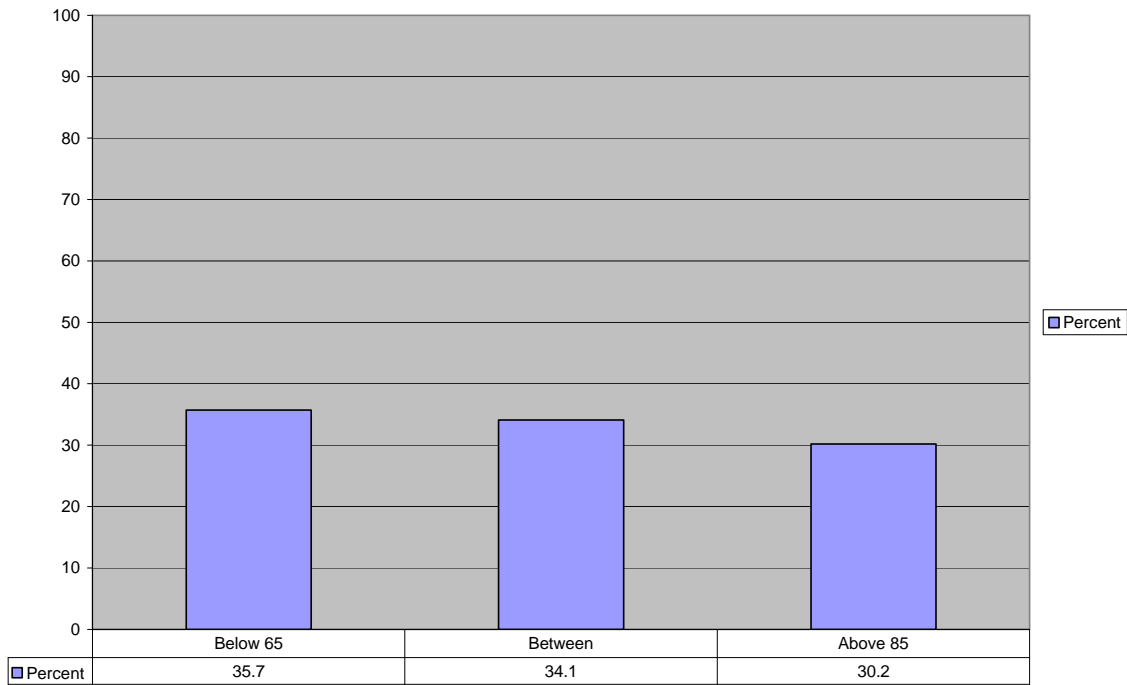


Figure 2. The percentage of students in each performance level using the cut scores after Round 2.

Percent in Category Round 3

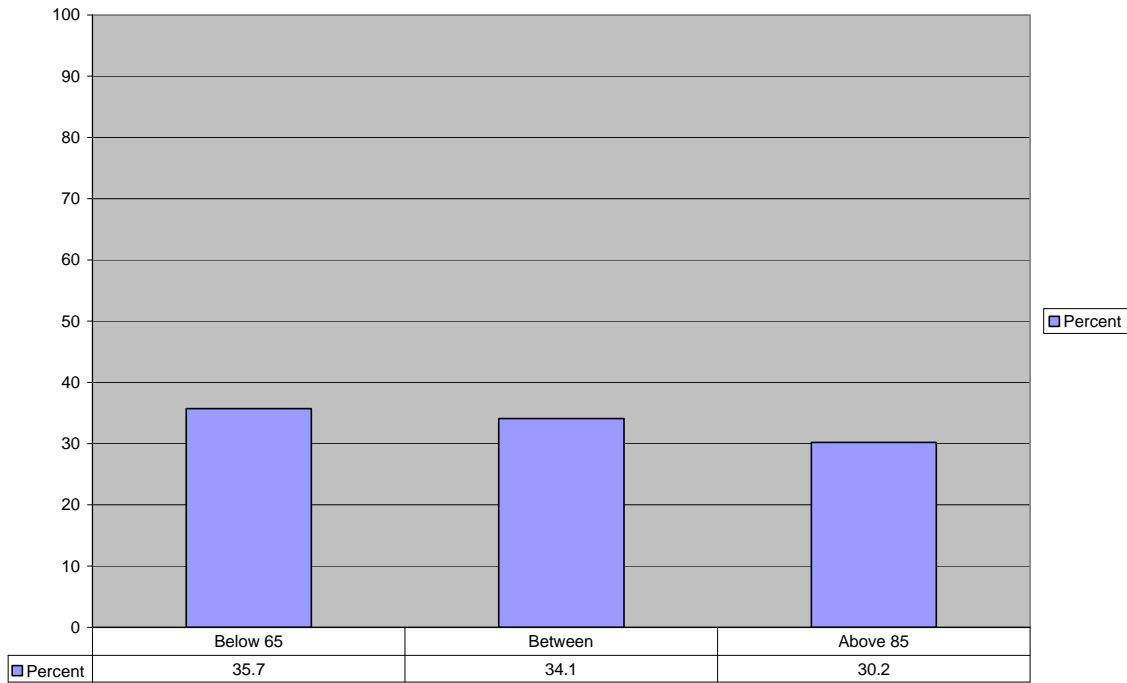


Figure 3. The percentage of students in each performance level using the cut scores after the Round 3 final rating.

Appendix A

Panelist Information Questionnaire for the New York State Regents Examination in Physical Setting/Earth Science Performance Test Standard Setting Meeting

**New York State Regents Examination in
Physical Setting/Earth Science Performance Test
Achievement Level Setting
Panelist Information Sheet**

Panelist ID: _____

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 Regents Examination in Physical Setting/Earth Science.

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)?

Large Medium Small

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

Urban Suburban Rural

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

North South East West Central

Appendix B
**Agenda for the New York State Regents Examination in Physical Setting/Earth
Science Performance Test Standard Setting Meeting**



**Recommendations for Setting Achievement Levels
for the New Earth Science Performance Test
Agenda**

DAY 1-September 27, 2007

Registration	8:30-9:00
Opening Remarks	9:00-9:30
Welcome and Why You Are Here	
Review of Agenda	
Security Forms	
Reimbursement	
Overview of Standard Setting	9:30-9:45
Purpose	
Item Mapping Methodology	
Overview of the Examination in Physical Setting/Earth Science	9:45-10:00
History	
Purposes	
Test Specifications	
BREAK	10:00-10:15
Complete performance test for Physical Setting/Earth Science	10:15-11:15
Introduce Achievement Level Descriptors	11:15-11:30
Construct Achievement Level Descriptors	11:30-12:00
Small Group Discussion	
LUNCH	12:00-1:00
(Train Table Leaders)	
Construct Achievement Level Descriptors	1:00-2:15
Small Group Discussion (continued)	
Large Group Discussion	

BREAK

2:15-2:30

Overview of Standard Setting 2:30-3:00
Item Mapping
Ordered Item Booklet
Item Map
Ratings Forms

BREAK 3:00-3:15

Practice Round 3:15-3:30

Round 1 Standard Setting 3:30-4:30
Readiness Form
Review Method
Collect page number/item numbers

End of Day Activities
Review Day 2 Schedule
Check in materials

END OF DAY 1

DAY 2 – September 28, 2007

Registration 8:00-8:30

Review schedule, answer questions 8:30-8:45

Feedback 8:45-9:15
Small group discussion of table agreement data

Round 2 Ratings 9:15-10:15
Readiness Form
Review Method
Collect page number/item numbers

BREAK 10:15-10:45

Feedback 10:45-11:30
Small group discussion of table agreement data
Large-group discussion of group agreement data
Large-group discussion of impact data

Round 3 Ratings 11:30-12:00

Readiness Form	
Review Method	
Collect page number/item numbers	
LUNCH	12:00-1:00
Feedback	1:00-1:30
Revisit Achievement Level Descriptors	1:30-3:00
Review items near recommended cut scores	
Revise achievement level descriptors	
Complete Survey	3:00-3:15
End of Day Activities	3:15-3:30
Check in materials	

Appendix C

Item Map for the New York State Regents Examination in Physical Setting/Earth Science Performance Test Standard Setting Meeting

Earth Science Item Map

Item in Binder	Item Location	Correct Answer	Standard	Key Idea
1	1-1	Allow 1 credit if only two or three mineral properties are correctly identified.		
2	a17-1	Allow 1 credit for identification of 1 plate movement in the correct sequence: A=Convergent, B=Divergent, C=Transform	4	2.1i
3	a14	1 (Allow 1 credit)	4	1.1d
4	a13	4 (Allow 1 credit)	4	1.1c
5	2	Allow 1 credit for the correct letter from the flowchart based upon student-identified properties.		
6	a17-2	Allow 2 credits for all 3 plate movements are in the correct sequence: A=Convergent, B=Divergent, C=Transform	4	2.1i
7	a2	4 (Allow 1 credit)	4	3.1c
8	20	Allow 1 credit for the correct eccentricity of the planet's orbit for the planet given on the station directions. Do <i>not</i> allow credit if a unit is given.	4	1.1b
9	11	Allow 1 credit if all three masses ($\pm 0.2g$) in the column labeled Mass of Cylinder with Fluid are correct and recorded to the <i>nearest tenth</i> of a gram.	1	MKI-1
10	a8	2 (Allow 1 credit)	4	3.1c
11	a5	4 (Allow 1 credit)	4	2.1g
12	a3	2 (Allow 1 credit)	4	2.2c
13	12	Allow 1 credit if the masses of all three fluids are correct as calculated based on the student's values for the mass of the cylinder with fluid and the mass of the empty cylinder. Mass must be recorded to the <i>nearest tenth</i> of a gram.	1	MKI-1
14	19	Allow 1 credit for the correct eccentricity of the contracted ellipse based on the student's answers for the distance between the foci and length of the major axis. The eccentricity must be expressed to the <i>nearest thousandth</i> . Do not allow credit if units, such as centimeters, are given.	4	1.2d
15	a9	1 (Allow 1 credit)	4	3.1c
16	a6	3 (Allow 1 credit)	4	2.2c
17	8	Allow 1 credit for the correct Station C time interval (± 10 seconds) . The time difference must appear in the table on the student answer page for credit.	4	2.1j
18	18	Allow 1 credit for the correct length of the major axis (± 1.0 cm) compared to the rater's determined length of the major axis for that code number. Students must record their answers to the <i>nearest tenth</i> of a centimeter.	4	1.1b
19	a1	1 (Allow 1 credit)	4	1.2c

20	a18	Allow 1 credit for the location East Pacific Ridge.	4	2.1n
21	7	Allow 1 credit for drawing the circle around Station B with the correct epicenter distance (± 200 km) . Measure the distance from the center of the dot at the two points where the student drawn circle intersects the predrawn circle. Allow this credit based on the numerical distance from Station B that was recorded in the student's answer booklet.	4	2.1j
22	5	Allow 1 credit for correctly classifying Rock Z.	4	3.1c
23	3	Allow 1 credit for correctly classifying Rock Y.	4	3.1c
24	a10	2 (Allow 1 credit)	4	1.2j
25	a4	2 (Allow 1 credit)	4	2.1g
26	a19	Allow 1 credit for correct placement and direction of arrow (see anchor paper).		2.1k
27	15	Allow 1 credit if the positions of all four fluids have been correctly labeled on the diagram based on the student's calculation of the densities.	1	MKI-3
28	a16	Allow 1 credit for the answer the dry and wet bulb temps are closer together.	4	2.1c
29	4	Allow 1 credit for providing a correct observable characteristic specific to Rock Y. If more then one rock characteristic is given, then all rock characteristics must be correct.	4	3.1c
30	10	Allow 1 credit for marking with an X the location of the epicenter at the place where the three circles intersect or nearly intersect. Allow no credit if more than one epicenter is indicated.	4	2.1j
31	a15	19°C (Allow 1 credit)	4	2.1c
32	6	Allow 1 credit for providing a correct observable characteristic specific to Rock Z. If more then one rock characteristic is given, then all rock characteristics must be correct.	4	3.1c
33	16	Allow 1 credit if the student has correctly identified the atmospheric temperature zone (layer) represented by fluid B based on where the student placed fluid B in the diagram.	4	2.1j
34	a12	2 (Allow 1 credit)	4	3.1c
35	14	Allow 1 credit for a correct calculation of all three densities based on the student's values for mass and volume. Densities must be recorded to the <i>nearest hundredth</i> of a gram per milliliter.	4	1.2c
36	17	Allow 1 credit if the center of the X is placed at the intersection of the constructed orbit and major axis (± 1 cm). The X must be closest to the focus labeled S . Do not allow credit if the student has not labeled one of the foci with an S or Sun.	4	2.1j
37	a11	2 (Allow 1 credit)	4	3.1c
38	13	Allow 1 credit if all three volumes (± 0.2 mL) are correct and recorded to the <i>nearest tenth</i> of a milliliter.	1	MKI-1
39	1-2	Allow 2 credits if all four mineral properties are correctly identified.	4	3.1a
40	a7	1 (Allow 1 credit)	4	2.1f

41	21	Allow 1 credit if the student marks the correct box and gives a correct explanation based on the student's eccentricity answers. Acceptable responses include, but are not limited to: The smaller eccentricity number indicates the more circular ellipse.	4	1.1b
42	9	Allow 1 credit for the Station C distance (± 200 km) based upon the student determined time difference. The distance must appear in the table on the student answer page for credit.	4	2.1j

Appendix D
**Readiness Survey for the New York State Regents Examination in Physical
Setting/Earth Science Performance Test Standard Setting Meeting**

Standard Setting Readiness Survey

Panelist ID: _____

Instructions: Please circle your response to the following questions.

Round 1		
I understand my task for Round 1.	No	Yes
I am ready to begin Round 1.	No	Yes

Round 2		
I understand my task for Round 2.	No	Yes
I understand the panelist agreement data that was presented from Round 1.	No	Yes
I understand the item difficulty data that was presented from Round 1.	No	Yes
I am ready to begin Round 2.	No	Yes

Round 3		
I understand my task for Round 3.	No	Yes
I understand the impact data that was presented from Round 2.	No	Yes

I am ready to begin Round 3.	No	Yes
------------------------------	----	-----

Appendix E

Questionnaire for the New York State Regents Examination in Physical Setting/Earth Science Performance Test Standard Setting

EVALUATION OF NEW YORK STATE REGENTS EXAMINATION IN PHYSICAL SETTING/EARTH SCIENCE PERFORMANCE TEST STANDARD SETTING

Please rate the following questions on a scale from 1 to 5, 1 being “**totally disagree**” and 5 being “**totally agree**”.

1. The method for setting standards, item mapping, was conceptually clear.

1 2 3 4 5

2. I had a good understanding of what the test was intended to measure.

1 2 3 4 5

3. I could clearly distinguish between student performance levels.

1 2 3 4 5

4. After the first round of ratings, I felt comfortable with the standard setting procedure.

1 2 3 4 5

5. I found the feedback on item difficulty useful in setting standards.

1 2 3 4 5

6. I found the feedback on the ratings of judges compared to other judges useful in setting standards.

1 2 3 4 5

7. I found the feedback on the percent of the students tested that would be classified at each performance level useful in setting standards.

1 2 3 4 5

8. I feel confident that the final cut score recommendations reflect the performance levels associated with the New York Regents Examination in Physical Setting/Earth Science Performance Test.

1 2 3 4 5

On the back of this page, please add any additional comment or observations on the standard setting process, facilitators, discussion, etc. Thank you.

Appendix F
Achievement Level Descriptions for New York State Regents Examination in
Physical Setting/Earth Science Performance Test

LEVEL I

- **Level I students can all measure with limited accuracy**
- **Level I students possess basic math skills**
- **Level I students follow simple directions**
- **Level I students can engage in concrete thinking**

LEVEL II

- **Level II students demonstrate proficient use of math skills and can solve equations**
- **Level II students can identify and interpret graphs, charts etc.**
- **Level II students follow multiple step directions**

LEVEL III

- **Level III students exhibit accuracy of measurement and math skills, etc.**
- **Level III students apply and draw conclusions based on concepts and skills and knowledge**
- **Level III students can do all of what Level I and Level II students can do**
- **Level III students have mastered Earth Science terms and concepts**

Appendix G

Earth Science Standard Setting Panelists

Name	School	City	State
Bobbi Jo Austin	Roslyn High School	Roslyn Heights	NY
Juan Betancourt	Rochester City School District	Rochester	NY
Rachel Colgan	Binghamton High School	Binghamton	NY
Teresa Doherty	Corcoran High School	Syracuse	NY
Kim Drake	Guilderland Central School	Guilderland	NY
Coralie Graham	Binghamton High School	Binghamton	NY
Louis Irizarry	South Jr. High School	Newburgh	NY
Christine Kola	Middle School 45	New York City	NY
Erichsen Kollmar	Groton High School	Groton	NY
Faye Landsman	Riverdale/Kingsbridge Academy MS/HS 141	Bronx	NY
Janette Liddle	Poland Central School District	Poland	NY
Michael McDonnell	Midwood High School at Brooklyn College	Brooklyn	NY
Brian McDowell	Hilton High School	Hilton	NY
Faye Melas	John Bowne High School	Flushing	NY
Barbara Merchant	Schenectady High School	Schenectady	NY
Joseph Perry	Palmyra Macedon Central High School	Palmyra	NY
John Pritchard	Grover Cleveland High School	New York City	NY
Thomas Rhindress	Croton-Harmon High School	Croton-On-Hudson	NY
Jesse Roehrich	Hospitality Management	New York City	NY
Dr. Frances Scelsi Hess	Cooperstown High School	Cooperstown	NY
Ashraf Shady	District 75/Citywide Programs	New York City	NY
Wendy Taylor	Schenectady City Schools	Schenectady	NY
Bernadette Tomaselli	Lancaster High School	Lancaster	NY
Gary Vorwald	Paul J. Gelinus Junior High School	Setaukat	NY
Ruth Wahl	Allegany-Limestone Central School	Allegany	NY
Esther Yee	Roslyn High School	Roslyn Heights	NY

Appendix H

Raw-score-to-theta distribution for Performance Part D test, produced using item parameters obtained from the field testing in May, 2006.

Raw Score	Theta
0	-3.712
1	-2.911
2	-2.110
3	-1.603
4	-1.219
5	-0.901
6	-0.624
7	-0.375
8	-0.143
9	0.076
10	0.288
11	0.496
12	0.703
13	0.914
14	1.130
15	1.357
16	1.600
17	1.867
18	2.173
19	2.539
20	3.022
21	3.789
22	4.556