New York State Regents Examination in Global History and Geography II

Standard Setting Technical Report

Prepared for the New York State Education Department
by Pearson

2019

Copyright

Developed and published under contract with the New York State Education Department by Pearson.

Copyright © 2019 by the New York State Education Department.

Secure Materials.
All rights reserved. No part of this document may be reproduced or transmitted by any means. Use of these materials is expressly limited to the New York State Education Department.

Table of Contents

Executive Summary 1

Regents Examination in Global History and Geography II 2

Performance Level Descriptions (PLDs) 5

Test Sample 7

Standard Setting 8

- Panelists 8
- Methodology 8
- Pre-workshop 9
- Workshop 9
- Pearson Standard Setting Website 10
- Test Review 10
- PLDs 10
- Modified Angoff Judgment Training 11
- Practice Judgment Task 12
- Standard Setting Rounds 12
- Level 2 Task Force 16
- Cut Scores and Impact Data 16
- Workshop Evaluation 18
- Final Recommendations 18

References 20

Appendix A: Standard Setting Workshop Agenda 21

Appendix B: Panelist Demographics 24

Appendix C: Panelist Readiness 28

Appendix D: Workshop Evaluation 30

Appendix E: Participant Evaluation Results 34

List of Tables

Table 1. Test Specifications for Regents Examination in Global History and Geography II: Multiple Choice 3
Table 2. Regents Examination in Global History and Geography II Design 4
Table 3. New York State Regents Examination Policy PLDs 6
Table 4. 2019 Test Sample for Standard Setting 7
Table 5. Geographic Locations of Panelists for Standard Setting Workshop 8
Table 6. Table Assignment of Global History and Geography II Claims 10
Table 7. Mock Table of Items with Greatest Variability in Judgments 14
Table 8. Feedback Data by Judgment Round 15
Table 9. Recommended Cut Scores Across Rounds 17
Table 10. Standard Error of Judgment 17
Table 11. Final Approved Cut Scores 19

List of Figures

Figure 1. Available Response Option to Judgment Question for MC and CRQ 11
Figure 2. Impact Data by Judgment Round 17
Figure 3. Impact Data based on Final Approved Cut Scores 19

Executive Summary

A standard setting meeting was conducted for the New York State Regents Examination in Global History and Geography II. The primary goal for this standard setting was to recommend cut scores that operationally define five performance levels: Level 1, Level 2, Level 3, Level 4, and Level 5. The performance level designations are used by local, state, and federal accountability programs and are central to communicating with parents, teachers, and the public. This document provides a detailed description of the activities held at the meeting.

The standard setting meeting was held June 17–18, 2019, in Albany, New York. Panelists were trained in and followed the Modified Angoff standard setting procedure, resulting in cut score recommendations that were brought to the New York State Education Department (NYSED).

In this report, panelists, materials, methodologies, and results are presented for the New York State Regents Examination in Global History and Geography II standard setting.

Regents Examination in Global History and Geography II

The Office of State Assessment at the New York State Department of Education worked with NYS social studies educators including the Social Studies Content Advisory Panel (CAP) to develop a high school Regents Examination in Global History and Geography II that measures the new Social Studies Framework adopted by the Board of Regents in 2014. This exam differs from the current Transition Examination in Global History and Geography – Grade 10 in the content measured (prior core curriculum versus new Framework).

The new Regents Examination in Global History and Geography II was administered for the first time in June 2019. The items on the assessment are designed to measure content and skills. Table 1 shows the range of items by Key Idea for the multiple-choice part of the assessment.

Table 1. Test Specifications for Regents Examination in Global History and Geography II: Multiple Choice

Key Idea	Range
10.1	0–3 0–11%
10.2	1–9 3–32%
10.3	1–9 3–32%
10.4	1–6 3–21%
10.5	1–9 3–32%
10.6	1–6 3–21%
10.7	1–9 3–32%
10.8	1-6 3-21%
10.9	1–9 3–32%
10.10	1–6 3–21%
Cross topical	0–6 0–21%
Total # of Multiple-Choice Items	28

The Regents Examination in Global History and Geography II contains three parts: part 1 with multiple-choice (MC) items, part 2 with constructed-response questions (CRQ), and part 3 the Enduring Issues Essay (EIE). The MC items are stimulus-based and require the students to select the correct response from four answer choices. Each MC item is worth one point. The CRQs are short-answer and designed to assess social science and historical thinking skills using primary and secondary sources. The CRQs are organized into two sets, and each set is based on a pair of documents that has three parts: context, sourcing, and relationship between documents. The CRQs are worth one point. Part 3 of the examination consists of an EIE question. The EIE question is based on five documents that each contain multiple enduring issues. The task requires students to identify and define an enduring issue based on a historically accurate interpretation of at least three documents, define the issue using evidence

from at least three documents, and argue that this is a significant issue that has endured. The argument should include how this issue has affected people or has been affected by people, and how the issue has continued to be an issue or has changed over time. The EIE is scored by two independent raters based on a five-point scale (0–5). The scores from each rater are averaged and then weighted by a factor of 3. The test design and score point distribution of each section is presented in Table 2.

Table 2. Regents Examination in Global History and Geography II Design

Parts	Question Type	Number of Questions	Maximum Raw Score	Weighting Factor	Maximum Weighted Score
Part 1	Stimulus-Based Multiple-Choice Questions	28	28	1	28
Part 2	Stimulus-Based Short-Answer Constructed-Response Questions (CRQs) - One Cause/Effect set - One Similarities/Differences or Turning Point set	Two Sets Set 1 has 3 one-point questions Set 2 has 4 one-point questions	7	1	7
Part 3	Enduring Issues Essay (EIE) - Extended Essay based on five documents	1	5	3	15
TOTAL					50

Performance Level Descriptions (PLDs)

PLDs are the foundation of standard setting activities because they provide the explanation of how student performance differs from one performance level to the next (Perie, 2008). In fact, PLDs are of such influence that, in a well-run standard setting workshop, they determine the rigor of the performance and thus the decisions made about placement of the cut score (Perie, Hess, and Gong, 2008). Moreover, PLDs serve multiple purposes in terms of communicating policy, facilitating test development, guiding standard setting, and providing score interpretation. Three types of PLDs (Egan, Schneider, and Ferrara, 2012) are used as an organizing framework for developing PLDs for the Regents Examination in Global History and Geography II:

- Policy PLD statements—Policy PLD statements are designed to capture the vision that an agency has for its performance levels. They specify the number of levels and the names for each level and summarize the expectations of student performance for a testing program, including any policy decisions being made at particular levels. Note that Table 3 provides the Policy PLDs for the New York State Regents Examination in Global History and Geography II.

- Range PLDs—Range PLDs are designed to describe the full range of performance for examinees at a given performance level. In other words, Range PLDs describe the aspects of test content or specific items that are indicative of a range of students at a specific performance level. Range PLDs can be informative in guiding item and test development as a testing program evolves. Range PLDs are also critical in that they are used to articulate a key component for standard setting, the Borderline Descriptions.

- Borderline Descriptions—Borderline Descriptions (also known as Threshold PLDs) are designed to articulate the transition points between the different ranges of performance defined by the Range PLDs. Specifically, Borderline Descriptions describe the knowledge and skills a student at the border between performance levels should know and be able to do. Because they articulate the specific performance that distinguishes levels of performance, Borderline Descriptions are typically used in standard setting activities. Range PLDs and Borderline Descriptions are clearly interdependent, which necessitates that they be developed in conjunction with each other.

Table 3. New York State Regents Examination Policy PLDs

Level 5: Students performing at Level 5 exceed the expectations of the Framework with distinction for Global History and Geography II.
Level 4: Students performing at Level 4 fully meet the expectations of the Framework for Global History and Geography II. They are likely prepared to succeed in the next level of coursework.
Level 3: Students performing at Level 3 minimally meet the expectations of the Framework for Global History and Geography II. They meet the content area requirements for a Regents diploma but may need additional support to succeed in the next level of coursework.
Level 2 (Safety Net): Students performing at Level 2 partially meet the expectations of the Framework for Global History and Geography II. Students with disabilities performing at this level meet the content area requirements for a local diploma but may need additional support to succeed in the next level of coursework.
Level 1: Students performing at Level 1 do not demonstrate sufficient knowledge, skills, and practices embodied by the Framework for Global History and Geography II for classification into a performance level.

Ultimately, the three types of PLDs are designed to describe the competencies of each performance level in relation to grade-level content standards while concurrently addressing their different functions. PLDs play a critical role in the standard setting process.

Test Sample

A sample of schools was identified to obtain test score data for setting performance standards on the Regents Examination in Global History and Geography II, given that the full set of operational data would not be available until the Fall. Note that schools had the option to administer either the Regents Examination in Global History and Geography II, the Transition Examination in Global History and Geography – Grade 10, or both. Schools were selected to form a representative sample in terms of demographic and achievement characteristics from data on the Transition Regents Examination in Global History and Geography – Grade 10 administered in June 2018. After the 2019 administration of the Regents Examination in Global History and Geography II, additional sampling adjustments were made to obtain a representative sample for standard setting purposes. The final sample consisted of approximately 11,500 students. Table 4 contains a summary of the standard setting sample and the 2018 population characteristics for comparison.

Table 4. 2019 Test Sample for Standard Setting

Demographics		Sample (2019)		Population (2018)	
		N	%	N	%
All Students		11,526	100.00	216,256	100.00
Gender	Female	5,768	50.04	107,496	49.71
	Male	5,758	49.96	108,760	50.29
Race/Ethnicity	American Indian or Alaska Native	111	0.96	1,366	0.63
	Asian	945	8.20	20,038	9.27
	Black or African American	1,964	17.04	39,995	18.49
	Hispanic or Latino	2,899	25.15	55,996	25.89
	Multiracial	186	1.61	3,510	1.62
	Native Hawaiian or Pacific Islander	31	0.27	578	0.27
	White	5,390	46.76	94,773	43.82
Students with Disabilities	No	9,711	84.25	183,772	84.98
	Yes	1,815	15.75	32,484	15.02
English Language Learner/Multilingual Learner	No	10,514	91.22	198,296	91.70
	Yes	1,012	8.78	17,960	8.30
Economically Disadvantaged	No	5,683	49.31	102,162	47.24
	Yes	5,843	50.69	114,094	52.76
Need Resource Capacity (NRC) Category	High Need: New York City	4,271	37.06	74,936	34.65
	High Need: Large Cities	652	5.66	8,084	3.74
	High Need: Urban/Suburban	789	6.85	15,674	7.25
	High Need: Rural	340	2.95	10,526	4.87
	Average Need	3,247	28.17	57,752	26.71
	Low Need	1,655	14.36	30,750	14.22
	Charter School	249	2.16	6,340	2.93
	Nonpublic School	323	2.80	12,194	5.64

Standard Setting

Cut scores for the Regents Examination in Global History and Geography II were recommended by a panel of 26 New York State educators over a two-day standard setting meeting. The Modified Angoff procedure (Angoff, 1971) of determining cut scores was used in a multi-round process of performance judgments, feedback data, and discussions.

Panelists

The panelists, recruited by NYSED, represented the major geographic regions of New York State, as shown in Table 5. A further breakdown of the demographics of the panelists can be found in Appendix B. There were 26 panelists who participated in standard setting. Sixteen panelists were female and 10 were male. Of the participants, 21 are social studies classroom teachers while the remaining participants hold positions as curriculum specialists, teacher leaders, or administrators.

Table 5. Geographic Locations of Panelists for Standard Setting Workshop

Geographic Location	Number of Panelists
Capital District	4
Central	4
Long Island	1
Lower Hudson	3
Mid-Hudson	1
Mid-West	1
North Country/Adirondacks	1
NYC	6
Southern Tier	1
Western	4

The entire group of panelists worked over the two days to recommend cut scores for Levels 3, 4, and 5. From these panelists, a subcommittee deemed the Level 2 Task Force was selected (total of 6 participants). The Level 2 Task Force was chosen to recommend cut scores for Level 2 separate from the full panel of panelists, given the characteristics of the student subgroup that mostly represents this level of performance. The Task Force convened at the end of Day 1 to recommend a cut score to be used to separate Level 1 from Level 2. This Task Force and its task will be discussed in more detail later in this report.

Methodology

The Modified Angoff standard setting procedure was used for this standard setting workshop. Panelists provided estimates of student success on each item of the Regents Examination in Global History and Geography II for each performance level. For each test item, panelists provided an estimate of the probability a student with performance at the borderline of each level would answer the question correctly. The probability was expressed as a percentage and explained as the likelihood a student with the knowledge and skills at the borderline of the performance level would get the item correct. For the MC items and CRQs, judgments were the percent chance students at a given performance level (in increments of 5 percentage points) would answer each item

correctly. For the EIE, panelists provided the score point that students at a given performance level would likely obtain.

The standard setting process, though, focuses on students *just barely* at each performance level, or *threshold* students. Therefore, the judgments provided by the panelists for each item and performance level were considered in terms of the success of threshold students. For example, *what is the probability that a student with performance at the borderline of each level would answer the question correctly* or *how many points would a student with performance at the borderline of each level likely earn if they answered the question?*

Pre-workshop

To engage in the judgment process of standard setting, there must be an understanding of content expectations for each performance level. Prior to the standard setting workshop, panelists were provided some pre-workshop tasks through the Pearson Standard Setting website: an introductory standard setting training video, review of sample Global History and Geography II items, and a review of the Policy and Range PLDs. These tasks were provided ahead of the workshop in order to set the context for standard setting. Panelists were also asked to review the Educator Guide that includes some sample test items—items available to the public as practice items—to understand some of what students had to do on the test.

The Policy and Range PLDs were provided so that panelists could review and understand the expectations within Global History and Geography II across state performance levels. Panelists were asked to review the Range PLDs in a survey format and take notes on their understanding of the Range PLDs using the following guiding questions:

- In what ways do the expectations increase from lower performance levels to higher performance levels?
- Within a performance level, are there any statements that differentiate achievement within the performance level? For example, high end of the performance level versus low end of the performance level?
- How different is student performance at the very bottom of a performance level compared to a student at the top of the previous performance level (i.e., lowest performing of Level 4 and highest performing of Level 3)?

The goal of this activity was for panelists to arrive at the workshop having a deeper understanding of the Range PLDs in order to better facilitate the development of Borderline Descriptions at the on-site meeting.

Workshop

The standard setting workshop was held in Albany, New York, June 17–18, 2019. Appendix A contains the workshop agenda. The workshop began with a welcome from NYSED, introductory remarks about the Regents Examination program, and the goals for setting performance standards on Global History and Geography II. A Pearson facilitator provided an overview of the standard setting process, explaining the different types of contextual information used (i.e., performance level descriptions, test content), the standard setting judgment process, and the different types of feedback data that

would be presented throughout the workshop. After the general orientation, including workshop logistics, the panelists began the workshop by reviewing the operational Regents Examination in Global History and Geography II.

Pearson Standard Setting Website

The Pearson Standard Setting website (Moodle) was used as the online platform for meeting pre-work facilitating the standard setting meeting, and for collecting panelist judgments throughout the standard setting process. Each panelist was provided a unique user identification and password that provide secure access to the site. Panelist access was restricted to the section of the site associated with the Global History and Geography II standard setting meeting.

The standard setting website provided panelists the opportunity to access all resource materials within a secure environment. Additionally, the website allowed for streamlining of the data collection from the individual judgment process.

Test Review

The panelists were provided the June 2019 test booklets that included the full operational test. This provided them with an opportunity to review the MC items, CRQs, and EIE to better understand what students were asked to do on the exam. The Rating Guide was provided via the standard setting website to provide the key idea assessed for each MC item, answer key for the MC items, scoring rubric for each CRQ, and exemplars and rubric for the EIE. A short discussion followed this task.

PLDs

After the test review, the facilitator discussed the Range PLDs and their use during the standard setting process. Panelists were given 15 minutes to discuss the Range PLDs in their table groups, focusing on key differences between the performance levels. The facilitator then provided an explanation for how to derive Borderline Descriptions from the Range PLDs. Prior to the standard setting, a set of draft Borderline Descriptions for Level 3, Claim 1 were drafted and presented to the panelists as part of a modeling activity. The facilitator walked the panelists through the guiding questions and illustrated how each descriptor was modified/constrained to create the drafts.

Following the modeling activity, the participants worked in their table groups to draft Borderline Descriptions for their assigned claim by accessing a Google doc through the website. Table 6 shows the claim that each table was assigned.

Table 6. Table Assignment of Global History and Geography II Claims

Table	Claim
1	Claim 1
2	Claim 2
3	Claim 3
4	Claim 4
5	Claim 1

After the panelists drafted Borderline Descriptions within their table, the facilitator organized the draft Borderline Descriptions from each table group into a master Google doc. The facilitator then led the whole group through a review of the descriptions and captured any group-approved edits into the master document. The Borderline Descriptions were printed and shared with the panelists to reference during the judgment activities.

Modified Angoff Judgment Training

The panelists were provided thorough training on how to make their recommendations as part of the standard setting meeting. They were instructed on using the Modified Angoff method (Angoff, 1971).

For each MC item and CRQ, participants were asked to answer the question:

"*What is the probability that a student with performance at the borderline of each level would answer the question correctly*?"

Significant time was spent on describing the thought process the panelists should go through using parts of the question.

- "What is the probability…"—Panelists will select an option that represents a range which contains an expected likelihood.
- "...that a student with performance at the borderline of each level…"—The panelists should reference the borderline PLDs for each performance level to determine how a student with performance at the borderline would be expected to respond.
- "...would…"—When considering expected student response to an item, the panelists needed to consider how a student would respond rather than how they should respond. Where "should" is an aspirational expectation, "would" is a more realistic expectation of a student response to the item.
- "...answer the question correctly?"—The panelists will review the knowledge, skills, and abilities necessary to provide a correct response to the item compared to the expected PLDs for the borderline performance level student.

Panelists were then instructed to ask the judgment question of how well students would do for each item for each performance level using the thought process they were trained on. Instead of having panelists provide open responses as expected probabilities of student success between 0% and 100%, they selected an option from 0% to 100% in intervals of 5%. A sample of the response options available to the panelists in the judgment survey is shown in Figure 1.

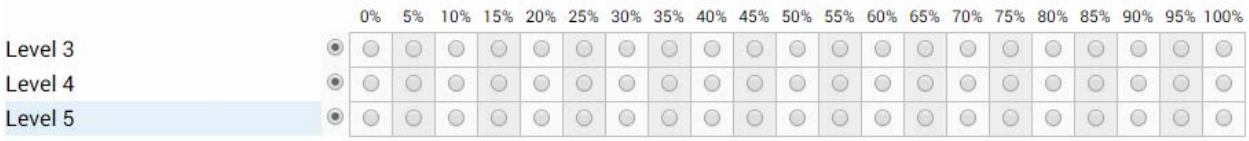


Figure 1. Available Response Option to Judgment Question for MC and CRQ

For the multi-point EIE item, participants were asked to answer the question:

"How many points would a student with performance at the borderline of each performance level likely earn if he or she answered the question?"

Significant time was spent on describing the thought process the panelists should go through using parts of the question.

- "How many points…"—Rather than recording the percent of students, panelists recorded the number of points for an item.
- "...would…"—When considering expected student response to an item, the panelists needed to consider how a student would respond rather than how they should respond. Where "should" is an aspirational expectation, "would" is a more realistic expectation of a student response to the item.
- "...a student with performance at the borderline of each performance level…"—The panelists referenced the borderline descriptions for the performance level to determine how a borderline student would be expected to respond.
- "...likely earn if he or she answered the question?"—In this context, "likely" is defined as 2 out of 3 times, or 67%. To make this concrete for panelists, facilitators asked them to think about 3 students at the borderline of a performance level.

Practice Judgment Task

At the end of the training session, panelists were provided the opportunity to practice making judgments prior to beginning the actual judgment rounds. The goals of this activity were for panelists to:

- get a feel for the range of different types of items and student responses they would encounter during the judgment task;
- experience the process of reviewing and making judgments for different types of items; and
- build their confidence that they understand the task that they are being asked to complete.

A set of 8 practice items was selected from the Educator Guide for use in this activity. These items were provided to the participants on paper. Items were selected to ensure that all item types were covered in the activity.

Following the practice judgments, the facilitator showed item-level judgment results interactively through the standard setting website, including what percentage of panelists selected each percent or point value for each performance level. The group also had the opportunity to discuss each practice item and to hear different perspectives on why panelists selected different probabilities for the MC items and CRQs and different point values for the EIE.

Standard Setting Rounds

Once training was completed, panelists began the actual judgment rounds. Prior to starting each judgment round, panelists were asked a series of readiness questions (via

a survey in the website) to verify that they understood their task and were ready to begin.

- Do you understand your task for the item judgment activity?
- Are you ready to begin the item judgment activity?

Following the readiness survey, the facilitator reviewed the responses. If a panelist were to have responded “no” to either of the questions in the readiness survey, the facilitator would have provided additional training and support as needed to the panelist. Once the facilitator ensured all panelists were ready to proceed, panelists were asked to make judgments for the first item starting at the lowest performance level based on the borderline descriptions and the knowledge, skills, and abilities required by the item. The panelists then made judgments for the same item for the rest of the performance levels before proceeding to the next item. Judgments were recorded in the website using the Item Judgment Survey. They were also provided a paper judgment record form to keep a record of their judgments for each round. Once the panelists completed making judgments for all items, they submitted their judgments for analysis.

After all panelists completed each judgment activity, data analysts from Pearson collected the item judgments from each participant from the site, performed the necessary analysis of the data and created feedback data that was provided to the panelists.

After Round 1, the facilitator provided cut scores generated from the panelists’ item-level judgments. Each panelist saw his/her own cut score for each performance level as well as a summary of cut scores from the entire committee. Here, panelists could compare their own cut scores to those from the overall committee and consider if their cut scores matched their level of expectations. To guide table group conversations, panelists received a list of flagged items per performance level that represented the items with the most disagreement in panelists’ judgments. In other words, the panelists were given lists of MC items and of CRQs that had the largest standard deviation in relation to the panelists’ judgments for each performance level. The lists were limited to eight items for each performance level. Table 7 is a mock table of items with the greatest variability in judgments across performance levels.

Panelists used this data to discuss differences in their judgments. The discussions throughout the workshop were meant to share perspectives in expectations, not to reach consensus on judgments.

Table 7. Mock Table of Items with Greatest Variability in Judgments

Global History and Geography II Round 1 Level L3 Flagged Items

UIN	Max Points	0	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
28MC	1	4%	.	.	.	4%	12%	4%	19%	12%	.	19%	4%	8%	8%	.	4%	.	.	4%	.	.
27MC	1	.	.	.	.	4%	4%	8%	8%	15%	4%	15%	4%	15%	12%	4%	4%	.	.	.	.	4%
29CRQ	1	.	.	.	.	4%	8%	12%	8%	12%	8%	23%	8%	8%	.	.	4%	4%	.	.	4%	.
21MC	1	.	.	8%	.	4%	15%	12%	15%	12%	8%	8%	4%	.	8%	.	8%	.	.	.	.	.
14MC	1	.	.	.	.	4%	4%	.	.	4%	4%	8%	19%	15%	8%	15%	.	15%	.	.	.	4%
4MC	1	.	.	.	.	.	8%	4%	4%	8%	8%	31%	8%	8%	4%	4%	8%	4%	.	.	.	4%
18MC	1	.	4%	.	.	4%	12%	12%	15%	15%	8%	12%	.	4%	.	12%	4%	.	.	.	.	.
25MC	1	.	.	.	.	8%	8%	8%	4%	4%	12%	15%	8%	15%	.	12%	8%	.	.	.	.	.

Global History and Geography II Round 1 Level L4 Flagged Items

UIN	Max Points	0	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
28MC	1	.	4%	.	.	.	4%	4%	4%	4%	.	8%	.	23%	.	12%	23%	8%	4%	.	.	4%
23MC	1	.	.	4%	.	4%	.	4%	.	4%	15%	12%	4%	4%	4%	19%	15%	4%	8%	.	.	.
21MC	1	.	.	.	.	8%	.	8%	.	4%	4%	8%	4%	19%	15%	8%	8%	8%	4%	4%	.	.
22MC	1	.	.	.	.	4%	.	8%	.	.	4%	8%	8%	8%	12%	12%	12%	15%	8%	4%	.	.
27MC	1	.	.	.	.	.	.	4%	8%	.	4%	15%	.	12%	8%	.	23%	15%	4%	4%	.	4%
18MC	1	.	.	.	4%	.	.	.	.	12%	12%	8%	12%	12%	15%	4%	4%	8%	4%	4%	4%	.
19MC	1	.	4%	.	.	.	.	8%	4%	.	12%	19%	12%	4%	12%	12%	8%	4%	4%	.	.	.
26MC	1	.	.	.	4%	.	.	4%	8%	4%	4%	12%	4%	19%	4%	15%	12%	8%	.	4%	.	.

Global History and Geography II Round 1 Level L5 Flagged Items

UIN	Max Points	0	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
28MC	1	.	.	4%	.	.	.	4%	4%	.	8%	.	.	.	4%	4%	8%	12%	19%	23%	8%	4%
23MC	1	.	.	.	4%	.	4%	.	4%	.	.	12%	4%	.	12%	.	4%	12%	15%	19%	12%	.
21MC	1	.	.	.	.	.	4%	4%	8%	.	4%	.	.	.	4%	4%	15%	12%	19%	15%	12%	.
19MC	1	.	.	.	4%	.	.	.	8%	4%	.	.	.	12%	4%	8%	19%	12%	12%	12%	8%	.
24MC	1	.	.	4%	.	.	4%	.	.	.	.	12%	8%	12%	4%	12%	8%	8%	15%	15%	.	.
20MC	1	.	.	.	.	.	4%	4%	4%	.	.	4%	.	4%	.	15%	8%	15%	23%	15%	4%	.
22MC	1	.	.	.	.	.	.	.	12%	.	.	4%	.	4%	4%	.	4%	19%	19%	19%	15%	.
5MC	1	.	.	.	.	.	.	.	12%	.	.	4%	4%	8%	12%	.	15%	12%	4%	23%	8%	.

Additionally, panelists were provided with empirical item difficulty data. These data reflected how well students actually did on each item from the operational test administration. Percent correct or mean score values were calculated from a representative sample of students who participated in the June 2019 test administration. These data were provided during the process to inform the panelists of their own content expectations. These data, though, reflect all levels of achievement. In other

words, the empirical item difficulty data were not just on the threshold, or *just barely*, students, but on all students sampled.

Round 2 of standard setting was performed just as Round 1 had been. The difference between the two rounds was that panelists were given feedback data and engaged in discussions prior to making Round 2 judgments. Panelists were instructed to revisit their judgments from Round 1 and make a new set of judgments, keeping their judgments from Round 1 or making revisions as they felt necessary. After Round 2 judgments, panelists were provided with another set of individual and committee-level cut score information as well as the item judgment variability data (e.g., Table 9) to discuss. The panelists discussed these data within their tables, but the facilitator led a larger group discussion on items with largest variability across the entire panel.

The facilitator also displayed impact data, or the distribution of students among performance levels based on the panel's overall cut scores. Presenting these data during the standard setting process gave the panelists the opportunity to see the consequences of their judgments and whether these consequences fit their expectations. The panelists were reminded that the data should not drive their judgments; rather, their judgments should be driven by content expectations. A discussion was led by the facilitator to discuss whether the impact data aligned with their content expectations.

Following the discussion of the Round 2 feedback data, the panelists provided one final round of judgments. This round was performed just as the previous two rounds. Once the results for Round 3 were complete, panelists were shown the final recommended cut scores and corresponding impact data. As a final task, the panelists completed a workshop evaluation that asked questions ranging from how comfortable they were with specific workshop activities to how comfortable they were with the final recommended cut scores. The workshop evaluation survey is provided in Appendix D. Table 8 displays the types of feedback data and at what round they were provided to the panelist.

Table 8. Feedback Data by Judgment Round

		Round 1	Round 2	Round 3
Item Level Feedback	Panelist Agreement Data	✓	✓	
	Item Means	✓		
	Score Point Distributions	✓		
Test Level Feedback	Individual Threshold Score	✓	✓	
	Table Threshold Score	✓	✓	
	Committee Threshold Score	✓	✓	✓

	Panelist Agreement Data	✓	✓	
	Impact Data		✓	✓

- Information about panelist cut scores for each performance level:
    - Individual cut scores: Item judgments for each performance level, recommended by each panelist, were summed across the items to obtain a cut score for each performance level. The cut score represents the minimum score needed to be classified into a performance level. The panelists were presented with their recommended cut score for each level, along with their recorded judgments for each item and each level as captured in the website survey. Panelists were asked to compare this output to their paper judgment record sheet to ensure what they expected to enter into the standard setting website was what was captured.
    - Committee cut score recommendations and statistics: Committee-level cut score recommendations were the median cut score across all panelists for each performance level. The committee members were presented with the committee-level recommendation and cut score statistics (minimum, maximum, median, mean, and standard deviation) for each level.
    - Panelist agreement data: Bar graphs showing the frequency of individual recommended cut scores for each performance level and across adjacent performance levels.
- Item-level judgment agreement across panelists: Distribution of individual item judgments for each item and performance level.
- Item means (p-values) and score point distributions: The average score earned by students for each item and the distribution of score points, for polytomously scored items, will be calculated from operational test data
- Impact data: Estimated proportion of students that would be classified into each performance level, based on the current recommended performance level cut scores, reflecting the performance of students who responded to the items during the June 2019 administration.

Level 2 Task Force

Cut scores for Level 2 were recommended by a representative group of five panelists from the larger standard setting panel. This group met at the end of the first day of standard setting to engage in an abbreviated process of recommending cut scores. The panelists discussed threshold descriptions for Level 2 and then provided one round of judgments, using the same process as was done for the other levels earlier in the day. The recommended Level 2 cut score and subsequent standard error of judgment was provided to NYSED for further deliberation.

Cut Scores and Impact Data

Cut scores were generated after each round of judgments. The median value of the individual panelists' cut scores, per performance level, was used as the recommended

cut score of the standard setting panel. Table 9 provides a summary of the cut scores for all three rounds.

Table 910. Recommended Cut Scores Across Rounds

Round 1			Round 2			Round 3		
Level 3	Level 4	Level 5	Level 3	Level 4	Level 5	Level 3	Level 4	Level 5
21.5	33.5	41.5	21.5	31.0	39.0	23.0	33.0	40.0

Additionally, the standard error of judgment (SEJ) was calculated for the final recommended cut scores to serve as additional information. The recommended performance level cuts plus and minus 3 standard errors of judgment are shown in Table 10.

Table 11. Standard Error of Judgment

Performance Level	SEJ	Cut – 3SEJ	Cut – 2SEJ	Cut – 1SEJ	Cut	Cut +1SEJ	Cut +2SEJ	Cut +3SEJ
Level 3	0.6	21.2	21.8	22.4	23.0	23.6	24.2	24.8
Level 4	0.7	30.9	31.6	32.3	33.0	33.7	34.4	35.1
Level 5	0.7	37.9	38.6	39.3	40.0	40.7	14.4	42.1

The impact data for all three rounds of standard setting are presented in Figure 2.

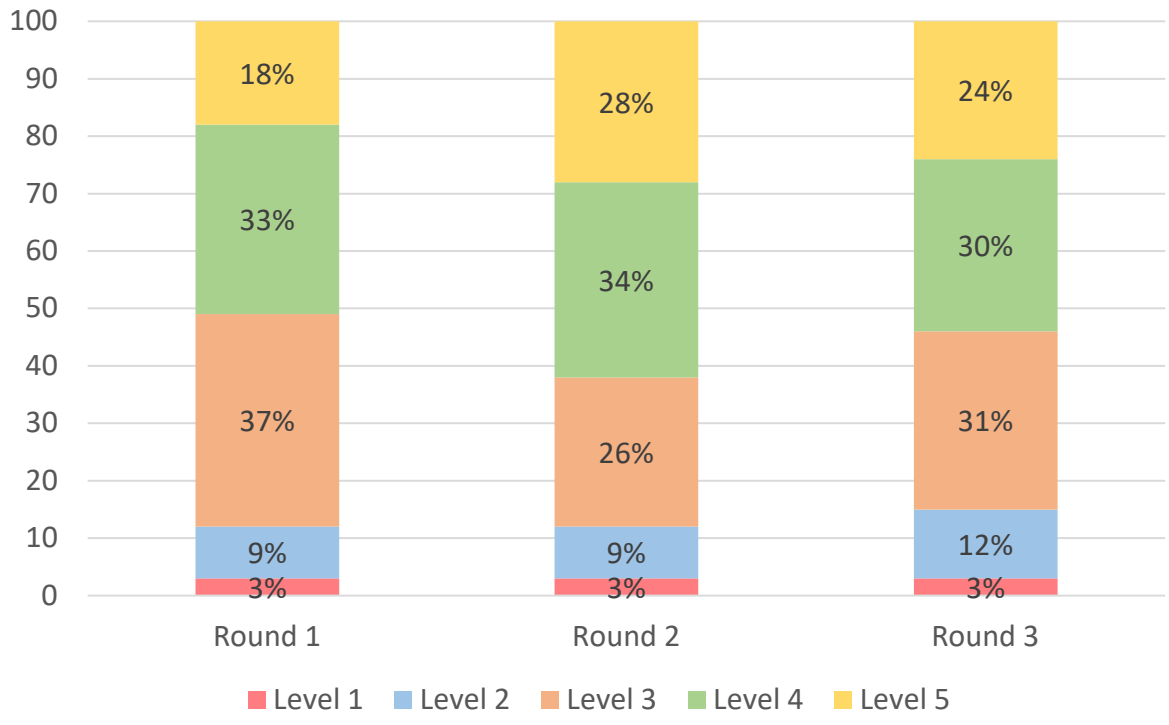


Figure 2. Impact Data by Judgment Round

Workshop Evaluation

Once the standard setting process was complete and the final recommended cut scores and impact data were shown, panelists completed a workshop evaluation on the various materials and activities of the standard setting process and the final recommended cut scores. The intent of this survey was to gather how well panelists understood the process, the materials used, and how comfortable they felt about the final recommended cut scores. For the survey items on the recommended cut scores, panelists were able to express how they would modify the percent of students classified into each performance level if they were somewhat uncomfortable with the overall final recommendation. Most items on this survey used a Likert scale, with different scales of affect (e.g., not confident to very confident, not adequate to very adequate, very useful to not useful, etc.) across the evaluation.

One particular evaluation item assessed panelists' confidence in the final cut scores: *How confident do you feel that the final cut score recommendations for Global History and Geography represent appropriate levels of student performance?* Of the 26 panelists, 21 marked "Confident" or "Very Confident" to this statement regarding the Level 3 cut, 24 for the Level 4 cut, and 24 for the Level 5 cut. There were no panelists that indicated they were "Not Confident" about the final recommended cut scores for any performance level. This highlights a crucial aspect of stakeholder involvement in this high stakes activity. The level of confidence expressed in regard to this particular item exhibits the support from the educators in the process of setting the performance standards for the Regents Examination in Global History and Geography II. In addition, some panelists provided handwritten quotes expressing their gratitude for being a part of this process and desire to participate in other educator panels for the New York State assessments. A complete summary of the evaluation results can be found in Appendix E.

Final Recommendations

As described in the previous sections, NYSED, with facilitation by Pearson, conducted a formal standard setting for the Regents Examination in Global History and Geography II. The goal of the standard setting meeting was to identify performance level cut scores consistent with the PLDs and state policy directives, using a standardized procedure called the Modified Angoff standard setting procedure.

The meeting reflected best practice as articulated in the Standards for Educational and Psychological Measurement (2014) and proceeded according to plans reviewed by the New York State Technical Advisory Committee. The panelists were diverse and representative of the State. The group followed, without incident, instructions delivered by the standard setting facilitator. All activities were formally overseen by the Office of State Assessment senior management and psychometric staff.

After careful consideration of the nature of the new examination, the rigor of the new curricula, the transitional and aspirational aspects of the New York State policy directives, and the role of the assessment in student learning throughout high school and beyond, the standard setting committee made recommendations on the cut scores to the Commissioner of Education. The Commissioner of Education subsequently made adjustments to the recommendations based on the committee feedback from the

survey, standard errors of judgment, and historical data. The final approved cut scores were implemented within the scale of measurement used to report student performance on the New York State Regent Examinations. The approved cuts scores are shown in Table 11 with subsequent impact data provided in Figure 3.

Table 1213. Final Approved Cut Scores

Level 2	Level 3	Level 4	Level 5
20	25.5	35	40

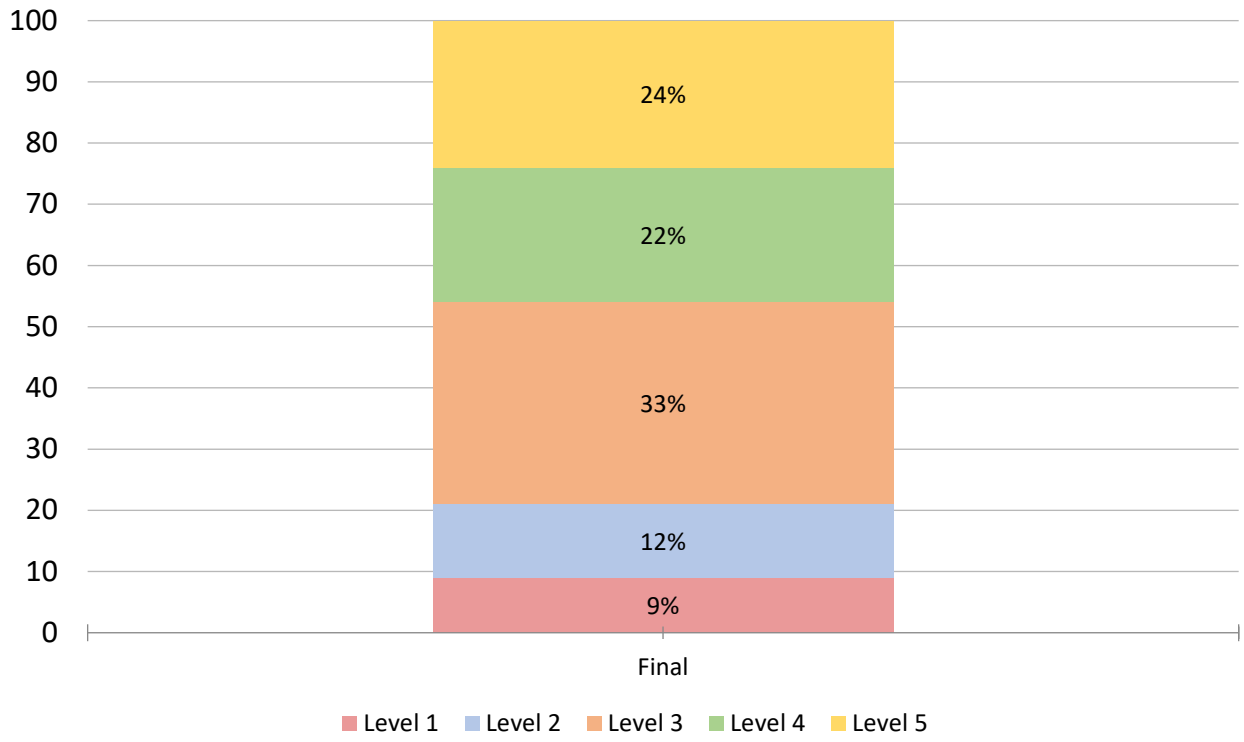


Figure 3. Impact Data based on Final Approved Cut Scores

References

American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education. Joint Technical Committee. (2014). *Standards for Educational and Psychological Testing.* Washington, DC: American Educational Research Association.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (pp.508–600). Washington, DC: American Council on Education.

Egan, K. L., Schneider, M. C., and Ferrara, S. (2012). Performance level descriptors: History, practice, and a proposed framework. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 79–106). New York, NY: Routledge.

Perie, M. (2008). A guide to understanding and developing performance-level descriptors. *Educational Measurement: Issues and Practice*, 27(4), 15–29.

Perie, M., Hess, K., and Gong, B. (2008). Writing performance level descriptors: Applying lessons learned from the general assessment to alternate assessments based on alternate and modified achievement standards. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York, NY, and available at http://www.nciea.org.

Appendix A: Standard Setting Workshop Agenda

University of the State of New York
State Education Department

Pearson

Regents Examination in Global History and Geography II
Standard Setting

Agenda

Day 1– June 17, 2019

7:30 – 8:00 a.m.	*Breakfast*
8:00 – 8:30 a.m.	Welcome and Standard Setting Overview
8:30 – 8:45 a.m.	Introductions, logins, material orientation, meeting security
8:45 – 10:15 a.m.	Experience the Assessment
10:15 – 10:30 a.m.	*Break*
10:30 – 11:00 a.m.	Review and Discuss Standards and Performance Level Descriptions
11:00 – 11:45 a.m.	Borderline Performance Level Descriptions Training
11:45 – 12:30 p.m.	*Lunch*
12:30 – 1:15 p.m.	Borderline PLD Table Discussion
1:15 – 2:15 p.m.	Borderline PLD Group Discussion
2:15 – 2:30 p.m.	*Break*
2:30 – 3:00 p.m.	Standard Setting Training
3:00 – 3:45 p.m.	Practice Judgment Activity and Discussion
3:45 – 5:00 p.m.	Round 1 Judgments
5:00 – 6:00 p.m.	*Break*
6:00 – 8:00 p.m.	Level 2 Task Force

Day 2– June 18, 2019

7:30 – 8:30 a.m.	*Breakfast*
8:30 – 8:45 a.m.	Round 1 Judgment Feedback
8:45 – 9:30 a.m.	Table Discussion – Round 1 Feedback
9:30 – 9:45 a.m.	Whole Group Discussion – Item Disagreement Data
9:45 – 10:45 a.m.	Round 2 Judgments
10:45 – 11:15 a.m.	*Break*
11:15 – 11:30 a.m.	Round 2 Judgment Feedback
11:30 – 12:00 p.m.	Table Discussion – Round 2 Feedback
12:00 – 12:45 p.m.	*Lunch*
12:45 – 1:30 p.m.	Whole Group Discussion – Round 2 Feedback
1:30 – 2:15 p.m.	Round 3 Judgments
2:15 – 2:45 p.m.	*Break*
2:45 – 3:15 p.m.	Round 3 Feedback, Evaluation, and Workshop Wrap-up

Appendix B: Panelist Demographics

While on-site at the standard setting meeting, panelists responded to an information survey in order to provide demographic and other pertinent information for validity evidence of the standard setting. A total of 26 panelists participated in the standard setting; however, only 24 panelists responded to the survey. The survey results have been tabulated below.

Current Position

Responses	Number of Committee Members
Classroom Teacher	21
Administrator	1
Other Position: Classroom teacher and curriculum specialist for 9–12	1
Other Position: Teacher Leader	1

Years of Professional Experience in Education

Responses	Number of Committee Members
1–5 years	5
6–10 years	3
11–15 years	4
16–20 years	5
More than 20 years	7

Years of Professional Experience in Teaching Global History

Responses	Number of Committee Members
1–5 years	8
6–10 years	3
11–15 years	5
16–20 years	5
More than 20 years	3

Experience with Special Populations

Responses	Number of Committee Members
Students receiving mainstream special education services	24
Students receiving self-contained special education services	6
Students who are English language learners	19
Students who are receiving general education instruction	23
Students who are receiving vocational technical instruction	6

Highest Degree Completed

Responses	Number of Committee Members
Master's degree (M.A., M.S.)	22
Doctoral degree (Ph.D., Ed.D.)	2

Gender

Responses	Number of Committee Members
Male	10
Female	14

Ethnicity

Responses	Number of Committee Members
Hispanic or Latino	3
Not Hispanic or Latino	19

Race

Responses	Number of Committee Members
Black or African American	5
White	16

Appendix C: Panelist Readiness

Panelist Readiness Survey Questions

Before beginning each round, panelists responded to the following questions via a survey in the standard setting website:

Practice Round and Round 1

Readiness Survey:

Before starting the activity, select a response for each of the following questions.

Do you understand your task for the Judgment activity?

Select one:

- Yes
- No

Are you ready to begin the Judgment activity?

Select one:

- Yes
- No

Rounds 2 and 3

Readiness Survey:

Before starting the activity, select a response for each of the following questions.

Do you understand your task for the Judgment activity?

Select one:

- Yes
- No

Do you understand the panelist feedback data that was presented?

Select one:

- Yes
- No

Are you ready to begin the Judgment activity?

Select one:

- Yes
- No

Appendix D: Workshop Evaluation

Regents Examination in Global History and Geography II
June 2019 Standard Setting

Evaluation

Panelists responded to the following evaluation questions via a survey in the standard setting website.

New York Regents Examination
Standard Setting Meeting

Global History and Geography II
Process Evaluation Survey

The purpose of this evaluation is to collect information about your experience with the activities of the standard setting meeting. Your opinions are an important part of our evaluation of this meeting.

Select the option that best reflects your opinion about the level of success of the various components of the standard setting meeting in which you participated. The activities were designed to help you both understand the process and be supportive of the recommendations made by the committee.

Select the option that best reflects your opinion about the level of success of the various components of the standard setting meeting in which you participated. The activities were designed to help you both understand the process and be supportive of the recommendations made by the committee.

	Not Successful	Partially Successful	Successful	Very Successful
General training on standard setting	○	○	○	○
Overview of the Global History and Geography II assessment	○	○	○	○
Experiencing the actual assessment	○	○	○	○
Discussion of the Range PLDs	○	○	○	○
Discussion and revision of the Borderline Descriptions	○	○	○	○
Overview of the standard-setting procedure	○	○	○	○
Practice exercise for the standard-setting procedure	○	○	○	○
Judgment rounds	○	○	○	○
Judgment round feedback - committee-level statistics	○	○	○	○
Judgment round feedback - panelist agreement data	○	○	○	○
Discussions after each round	○	○	○	○

How useful do you feel the following activities or information were in assisting you to make your recommendations?

	Very Useful	Useful	Somewhat Useful	Not Useful
Range Performance Level Descriptions (PLDs)	○	○	○	○
Borderline Descriptions	○	○	○	○
Committee-level statistics	○	○	○	○
Panelist agreement data provided after Round 1	○	○	○	○
Panelist agreement data provided after Round 2	○	○	○	○
Discussion after each judgment round	○	○	○	○

How adequate were the following elements of the session?

	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate
Training provided on the standard-setting process	○	○	○	○
Amount of time spent training	○	○	○	○
Total amount of time to review and discuss Borderline Descriptions	○	○	○	○
Total amount of time to discuss the practice judgments	○	○	○	○
Amount of time to make judgments	○	○	○	○
Visual presentation of the feedback provided	○	○	○	○
Number of judgment rounds	○	○	○	○

How confident do you feel that the final cut score recommendations for Global History and Geography II represent appropriate levels of student performance?

	Not Confident	Somewhat Confident	Confident	Very Confident
Level 3	○	○	○	○
Level 4	○	○	○	○
Level 5	○	○	○	○

I feel the percentage of students in Level 3 is:

	Too High	About Right	Too Low
Level 3	○	○	○

I would expect the percentage of students in Level 3 to be:

I feel the percentage of students in Level 4 is:

	Too High	About Right	Too Low
Level 4	○	○	○

I would expect the percentage of students in Level 4 to be:

I feel the percentage of students in Level 5 is:

	Too High	About Right	Too Low
Level 5	○	○	○

I would expect the percentage of students in Level 5 to be:

Answer the following questions about your overall experience at the standard setting meeting.

How adequate were the following elements of the standard setting session?

	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate
Facilities used for the standard setting	○	○	○	○
Computers used during the meetings	○	○	○	○
Standard Setting website for accessing materials and making judgments	○	○	○	○
Materials provided in the folder	○	○	○	○
Work space in table groups during meeting	○	○	○	○

Did you have adequate opportunities during the session to:

	Not Adequate	Somewhat Adequate	Adequate	More Than Adequate
Express your opinions about student performance levels	○	○	○	○
Ask question about the cut scores and how they will be used	○	○	○	○
Ask questions about the process of making cut score recommendations	○	○	○	○
Interact with you fellow panelists	○	○	○	○

Do you believe your opinions and judgments were treated with respect by:

	No	Sometimes	Yes
Fellow panelists	○	○	○
Facilitators	○	○	○

Please use the space below to provide any additional comments you have regarding the standard setting process, facilitators, materials, etc.

Paragraph B I

Path: p

Appendix E: Participant Evaluation Results

Question 1: Select the option that best reflects your opinion about the level of success of the various components of the standard setting meeting in which you participated. The activities were designed to help you both understand the process and be supportive of the recommendations made by the committee.

Responses	Not Successful	Partially Successful	Successful	Very Successful
General training on standard setting	0	0	13 (50%)	13 (50%)
Overview of the Regents Examination in Global History and Geography II	0	0	7 (27%)	19 (73%)
Experiencing the actual examination	0	0	10 (38%)	16 (62%)
Discussion of the Range PLDs	0	0	13 (50%)	13 (50%)
Discussion and revision of the Borderline Descriptions	0	1 (4%)	12 (46%)	13 (50%)
Overview of the standard setting procedure	0	1 (4%)	11 (42%)	14 (54%)
Practice exercise for the standard setting procedure	0	0	9 (35%)	17 (65%)
Judgment rounds	0	1 (4%)	10 (38%)	15 (58%)
Judgment round feedback - committee-level statistics	0	0	10 (38%)	16 (62%)
Judgment round feedback - panelist agreement data	0	1 (4%)	9 (35%)	16 (62%)
Discussions after each round	0	0	13 (50%)	13 (50%)

Question 2: How useful do you feel the following activities or information were in assisting you to make your recommendations?

Responses	Very Useful	Useful	Somewhat Useful	Not Useful
Range Performance Level Descriptions (PLDs)	11 (42%)	13 (50%)	2 (8%)	0
Range Performance Level Descriptions (PLDs)	15 (58%)	10 (38%)	1 (4%)	0
Committee-level statistics	15 (58%)	10 (38%)	1 (4%)	0
Panelist agreement data provided after Round 1	15 (58%)	11 (42%)	0	0
Panelist agreement data provided after Round 2	16 (62%)	10 (38%)	0	0
Discussion after each judgment round	16 (62%)	10 (38%)	0	0

Question 3: How adequate were the following elements of the session?

Responses	Not Adequate	Partially Adequate	Adequate	Very Adequate
Training provided on the standard-setting process	0	0	13 (50%)	13 (50%)
Amount of time spent training	0	1 (4%)	11 (42%)	14 (54%)
Total amount of time to review and discuss Borderline Descriptions	0	1 (4%)	10 (38%)	15 (58%)
Total amount of time to discuss the practice judgments	0	1 (4%)	10 (38%)	15 (58%)
Amount of time to make judgments	0	2 (8%)	9 (35%)	15 (58%)
Visual presentation of the feedback provided	0	0	10 (38%)	16 (62%)
Number of judgment rounds	0	0	13 (50%)	13 (50%)

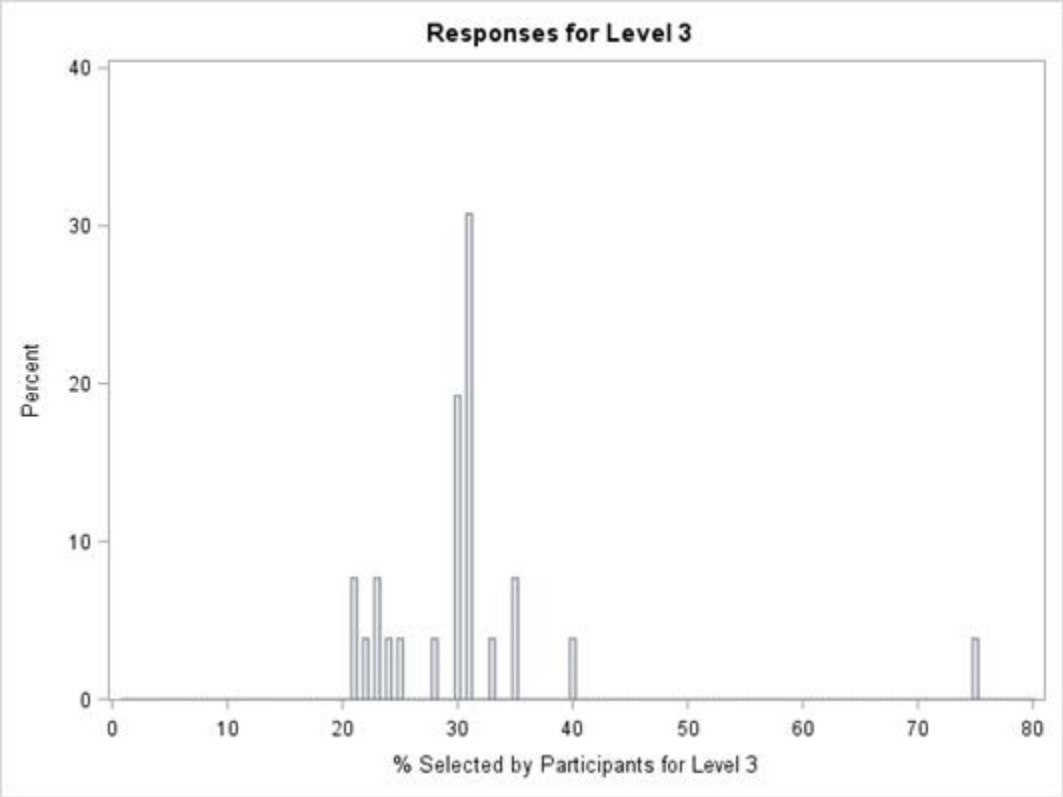
Question 4: How confident do you feel that the final cut score recommendations for Global History and Geography II represent appropriate levels of student performance?

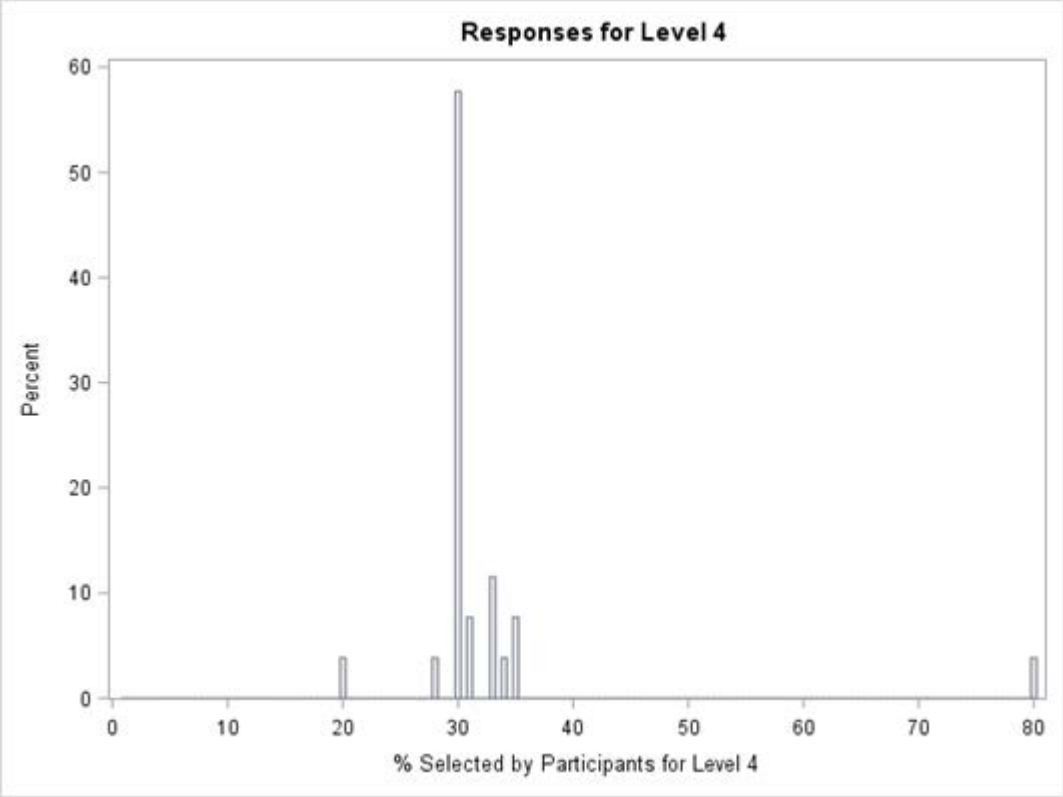
Responses	Not Confident	Somewhat Confident	Confident	Very Confident
Level 3	0	5 (19%)	7 (27%)	14 (54%)
Level 4	0	2 (8%)	9 (35%)	15 (58%)
Level 5	0	2 (8%)	9 (35%)	15 (58%)

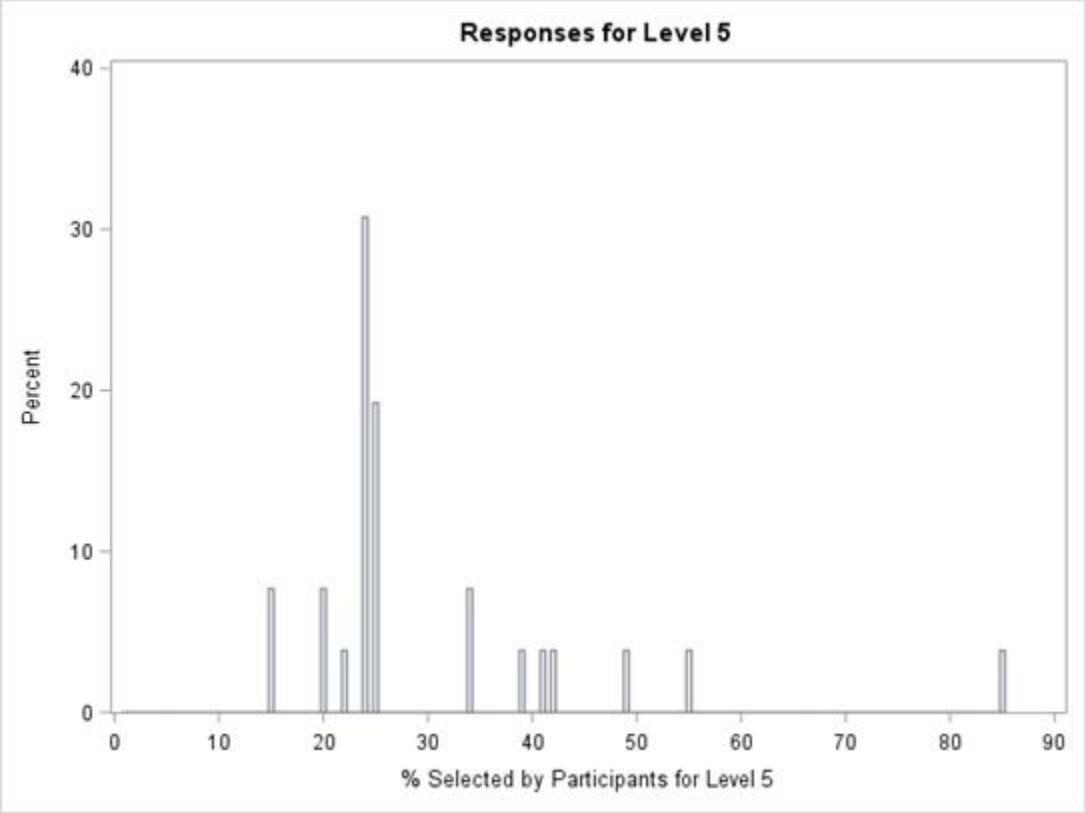
Question 5 (Level 3), Question 7 (Level 4), and Question 9 (Level 5): I feel the percentage of students in each level is:

Responses	Too High	About Right	Too Low
Level 3	0	26 (100%)	0
Level 4	0	25 (96%)	1 (4%)
Level 5	3 (12%)	22 (85%)	1 (4%)

Question 6 (Level 3), Question 8 (Level 4), and Question 10 (Level 5): I would expect the percentage of students in each level to be:






Responses for Level 5
Percent
40
30
20
10
0
0
10
20
30
40
50
60
70
80
90
% Selected by Participants for Level 5

Question 11: How adequate were the following elements of the session?

Responses	Not Adequate	Partially Adequate	Adequate	Very Adequate
Facilities used for the standard setting	0	0	12 (46%)	14 (54%)
Computers used during the meetings	0	0	9 (35%)	17 (65%)
Standard Setting website for accessing materials and making judgments	0	1 (4%)	6 (23%)	19 (73%)
Materials provided in the folder	0	0	8 (31%)	18 (69%)
Work space in table groups during meeting	2 (8%)	3 (12%)	8 (31%)	13 (50%)

Question 12: Did you have adequate opportunities during the session to:

Responses	Not Adequate	Partially Adequate	Adequate	Very Adequate
Express your opinions about student performance levels	0	1 (4%)	8 (31%)	17 (65%)
Ask questions about the cut scores and how they will be used	0	0	8 (31%)	18 (69%)
Ask questions about the process of making cut score recommendations	0	0	9 (35%)	17 (65%
Interact with your fellow panelists	0	0	5 (19%)	21 (81%)

Question 13: Do you believe your opinions and judgments were treated with respect by:

Responses	No	Sometimes	Yes
Fellow panelists	0	1 (4%)	25 (96%)
Facilitator	0	0	26 (100%)