New York State Regents Examination in United States History and Government (Framework)

Standard Setting Technical Report

Prepared for the New York State Education Department
by Pearson

2023

Copyright

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

Copyright © 2023 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

List of Tables

List of Figures

Executive Summary

A standard setting meeting was conducted for the New York State Regents Examination in United States History and Government (Framework). 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 13–14, 2023, 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 United States History and Government (Framework) standard setting.

Regents Examination in United States History and Government (Framework)

The Office of State Assessment (OSA) at the New York State Education Department (NYSED) worked with NYS social studies educators, including the Social Studies Content Advisory Panel (CAP), to develop a high school Regents Examination in United States History and Government that measures the new Social Studies Framework adopted by the Board of Regents in 2014. This exam differs from the previous Regents Examination in United States History and Government in the content measured (prior core curriculum vs. new Framework).

The new Regents Examination in United States History and Government (Framework) was administered for the first time in June 2023, with items designed to measure content and skills. The assessment contains three parts: Part 1 with multiple-choice items; Part 2 with two short-essay items, each based on a pair of documents; and Part 3 with short-response items and an extended essay item. Table 1 presents the composition of the Regents Examination in United States History and Government (Framework) by part along with the related scoring.

Table 1. Regents Examination in United States History and Government (Framework) Design

Part	Item Type	Number of Items	Maximum Raw Score	Weighting Factor	Maximum Weighted Score
Part 1	Stimulus-Based Multiple-Choice Items	28	28	1	28
Part 2	Stimulus-Based Short-Essay Items: – Set 1: Students describe the historical context between two documents and identity and explain the relationship between the events and/or ideas found in those documents. (Cause/Effect **or** Similarity/Difference **or** Turning Points) – Set 2: Student describe the historical context surrounding two documents **and** analyze and explain how *audience*, **or** *purpose,* **or** *point of view* affects the document's use as a reliable source.	**Two Sets:** Set 1 has one short-essay item based on a 5-point rubric. Set 2 has one short-essay item based on a 5-point rubric.	10	1	10
Part 3	Civic Literacy Document-Based Essay – Short-response items based on each of the six documents – Extended essay based on the set of six documents and focused on constitutional and civic issues	6 1	6 5	1 3	6 15
TOTAL					59

Part 3 of the examination includes an extended essay-based item based on six documents with the option of additional documents in an a/b format. The essay asks students to complete three tasks:

- Describe the historical circumstances surrounding this constitutional or civic issue.
- Explain efforts to address this constitutional or civic issue by individuals, groups, and/or governments.
- Either discuss the extent to which these efforts were successful **or** discuss the impact of these efforts on the United States and/or American society.

The essay is scored by two independent raters based on a 5-point scale (0–5). The scores from each rater are averaged and then weighted by a factor of 3. Table 2 presents the test design and score point distribution of each section, including the range of items by Key Idea for the multiple-choice part of the assessment.

Table 2. Regents Examination in United States History and Government (Framework) Test Specifications: Multiple Choice

	Range	
Key Idea	Number of Items	Percentage of Items
11.1	0–4	0–14%
11.2	2–5	7–18%
11.3	2–5	7–18%
11.4	0–4	0–14%
11.5	0–5	0–18%
11.6	0–4	0–14%
11.7	0–4	0–14%
11.8	0–4	0–14%
11.9	0–5	0–18%
11.10	0–5	0–18%
11.11	0–3	0–11%
Cross topical	0–5	0–18%
Total Number of Multiple-Choice Items	28	

Performance Level Descriptions (PLDs)

Performance level descriptions (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). 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 et al., 2008). PLDs also serve multiple purposes in terms of communicating policy, facilitating test development, guiding standard setting, and providing score interpretation. Three types of PLDs (Egan et al., 2012) are used as an organizing framework for developing PLDs for the Regents Examination in United States History and Government (Framework):

- Policy PLD statements are designed to capture the vision 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. Table 3 presents the Policy PLDs for the New York State Regents Examination in United States History and Government (Framework).
- Range PLDs are designed to describe the full range of performance for students 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. They are critical in that they are used to articulate the Borderline Descriptions, which are a key component for standard setting.
- 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, they 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 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 this level meet the expectations of the Framework with distinction for United States History and Government.
Level 4: Students performing at this level fully meet the expectations of the Framework for United States History and Government. They are likely prepared to succeed in the next level of coursework.
Level 3: Students performing at this level minimally meet the expectations of the Framework for United States History and Government. 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: Students performing at this level partially meet the expectations of the Framework for United States History and Government. Students with disabilities performing at this level meet the content are requirements for a local diploma but may need additional support to succeed in the next level of coursework.
Level 1: Students performing at this level demonstrate knowledge, skills, and practices embodied by the Framework for United States History and Government below that of Level 2.

Ultimately, the three types of PLDs are designed to describe the competencies of each performance level in relation to grade-level content standards while 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 United States History and Government (Framework), given that the full set of operational data would not be available until the fall. Schools were selected to form a representative sample in terms of demographic and achievement characteristics from data on the Regents Examination in United States History and Government administered in June 2019 – the last time the examination was administered. After the 2023 administration of the Regents Examination in United States History and Government (Framework), additional sampling adjustments were made to obtain a representative sample for standard setting purposes. Table 4 presents a summary of the standard setting sample and the 2019 population characteristics for comparison. The final sample consisted of approximately 12,000 students.

Table 4. 2023 Test Sample for Standard Setting

Demographics		Sample (2023)		Population (2019)	
		N	%	N	%
	All Students	12,088	100.00	202,139	100.00
Gender	Female	5,931	47.91	101,458	50.19
	Male	5991	50.64	100,681	49.81
Race/Ethnicity	American Indian or Alaska Native	418	3.5	1,319	0.65
	Asian	1,046	8.77	20,411	10.10
	Black or African American	1,897	15.9	34,466	17.05
	Hispanic or Latino	2,278	19.09	48,683	24.08
	Multiracial	270	2.26	3,143	1.55
	Native Hawaiian or Pacific Islander	53	0.44	565	0.28
	White	5,969	50.03	93,552	46.28
Students with Disabilities	No	10,385	85.91	175,757	86.95
	Yes	1,703	14.09	26,382	13.05
English Language Learner	No	11,367	94.04	189,091	93.55
	Yes	721	5.96	13,048	6.45
Economically Disadvantaged	No	6,425	53.15	103,191	51.05
	Yes	5,663	46.85	98,948	48.95
Need Resource Capacity (NRC) Category	High Need: New York City	3,150	26.06	68,362	33.82
	High Need: Large Cities	475	3.93	6,694	3.31
	High Need: Urban/Suburban	1,071	8.86	13,360	6.61
	High Need: Rural	719	5.95	9,550	4.87
	Average Need	3,826	31.65	54,741	27.08
	Low Need	1,890	15.64	29,563	14.63
	Charter School	213	1.76	6,188	3.06
	Nonpublic School	744	6.15	13,681	6.77

Note. The 2023 population is being compared to the 2019 population because 2019 was the last time the Regents Examination in United States History and Government was administered.

Standard Setting

Cut scores for the Regents Examination in United States History and Government (Framework) were recommended by a panel of 24 New York State (NYS) 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 NYS, as shown in Table 5. Appendix B presents a further breakdown of the panelist demographics. A total of 24 panelists participated in the standard setting, with 13 female panelists and 10 male panelists. One panelist chose not to provide information regarding their gender. Of the panelists, 23 were social studies classroom teachers while the remaining panelist identified as a school administrator. The participants were selected to represent experience with teaching the variation of students across the state.

Table 5. Geographic Locations of Standard Setting Panelists

Geographic Location	Number of Panelists
Capital District	4
Central	5
Long Island	3
Lower Hudson	1
Mid-Hudson	2
North Country/Adirondacks	1
NYC	4
Southern Tier	0
Western	4

The panel 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 (with a total of four participants). The Level 2 Task Force was chosen to recommend cut scores for Level 2 separate from the full panel given the characteristics of the student subgroup that mostly represents this level of performance. The Task Force convened at the end of Days 1 and 2 to recommend a cut score to be used to separate Level 1 from Level 2.

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 United States History and Government (Framework) for each performance level. For each item, panelists provided an estimate of the probability a student with performance at the borderline would answer the item 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 multiple-choice items, judgments were the percent chance students at a given performance level would answer each item correctly (in increments of 5 percentage points). For the short-response essays and extended essay, panelists provided the score point that students at a given performance level would likely obtain.

The standard setting process focused on students *just barely* at each performance level, or *threshold* (borderline) 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, including an introductory standard setting training video, review of sample United States History and Government (Framework) items, and a review of the Policy and Range PLDs. These tasks were provided ahead of the workshop 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 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 United States History and Government (Framework) across performance levels. Panelists were asked to review the Range PLDs in a survey format and take notes on their understanding of them 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 (e.g., high end of the performance level vs. 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 (e.g., lowest performing of Level 4 vs. 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 to better facilitate the development of the Borderline Descriptions.

Workshop

The standard setting workshop was held in Albany, New York, from June 13–14, 2023. Appendix A presents 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 United States History and Government (Framework). A Pearson facilitator provided an overview of the standard setting process, explaining the different types of contextual information used (e.g., PLDs, 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 United States History and Government (Framework).

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 collecting panelist judgments throughout the standard setting process. Each panelist was provided a unique user identification and password that provided secure access to the site. Panelist access was restricted to the section of the site associated with the United States History and Government (Framework) standard setting meeting. The standard setting website provided panelists the opportunity to access all resource materials within a secure environment. The website also allowed for streamlining of the data collection from the individual judgment process.

Test Review

The panelists were provided the June 2023 test booklets that included the full operational test. This provided them with an opportunity to review the multiple-choice items, short-essay items, short-response items, and the extended essay 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 multiple-choice item, the answer key for the multiple-choice items, the scoring rubric for each short-essay item, and exemplars and the rubric for the extended essay item. 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 panelists worked in their table groups to draft Borderline Descriptions for their assigned claim by accessing a Google doc through the website. Table 6 presents the claim that each table was assigned.

Table 6. Table Assignment of United States History and Government (Framework) Claims

Table	Claim
1	Claim 1
2	Claim 2
3	Claim 3
4	Claim 4a
5	Claim 4b

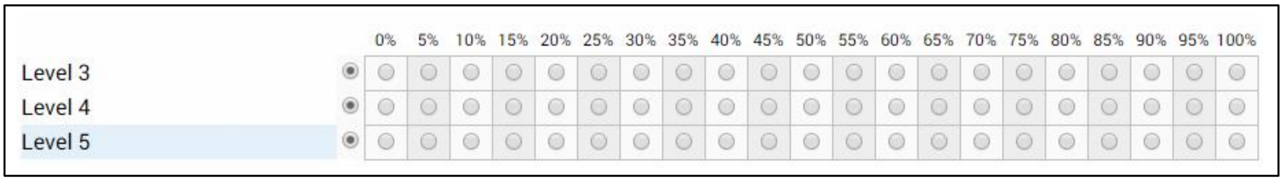
After the panelists drafted the Borderline Descriptions within their table, the facilitator organized the draft descriptions from each table group into a master Google doc. The facilitator then led the whole group through a review of the descriptions and captured any group-approved edits into the master 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 multiple-choice item, panelists 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 that contains an expected likelihood.
- "...that a student with performance at the borderline of each level…"—Panelists should reference the Borderline Descriptions for each performance level to determine how a student with performance at the borderline would be expected to respond.
- "...would…"—When considering the 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 answer the judgment question 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%. Figure 1 presents a sample of the response options available to the panelists in the judgment survey.



		0%	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%	55%	60%	65%	70%	75%	80%	85%	90%	95%	100%
Level 3	●	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Level 4	●	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Level 5	●	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○

Figure 1. Available Response Options to Judgment Question for Multiple-Choice Items

For the short-response essays and the extended essay, panelists 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 the 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 three 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 item types, and
- build their confidence that they understand the task they are being asked to complete.

A set of eight practice items was selected from the Educator Guide for use in this activity and provided to the panelists 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 multiple-choice items and different point values for the short-response essays and the extended essay.

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 on the website, as shown in Appendix C) to verify that they understood their task and were ready to begin:

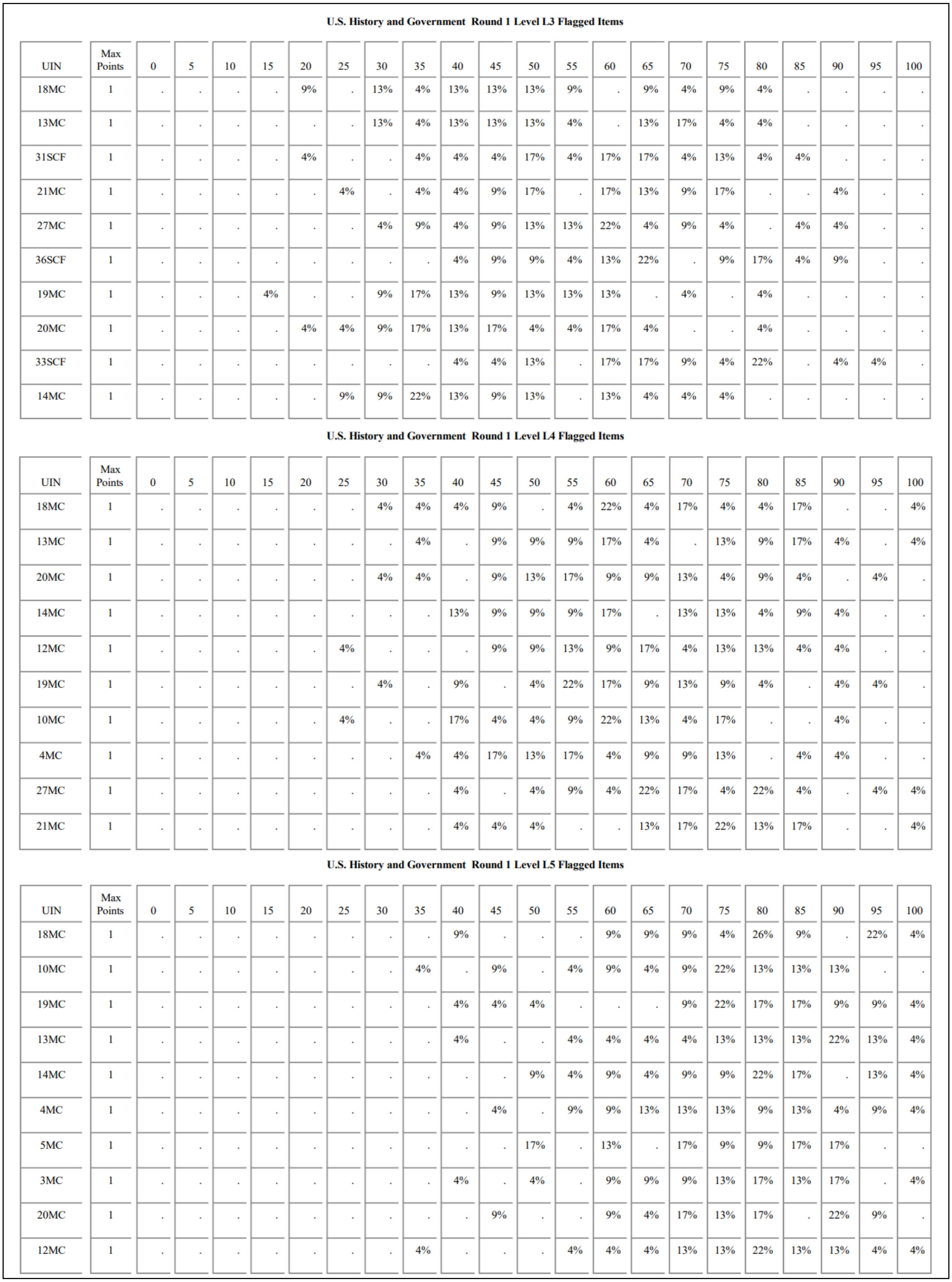
- Do you understand your task for the item judgment activity?
- Are you ready to begin the item judgment activity?

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 that 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 on 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, Pearson data analysts collected the item judgments, performed the necessary analysis of the data, and created feedback data that were provided to the panelists.

After Round 1, the facilitator provided cut scores generated from the panelists' item-level judgments. Each panelist saw their own cut score for each performance level and a summary of cut scores from the entire panel. Here, panelists could compare their own cut scores to those from the overall panel 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 most disagreement in panelists' judgments (i.e., the panelists were given lists of multiple-choice items 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. Figure 2 presents 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.

Panelists were also provided with empirical item difficulty data that reflected how well students performed 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 2023 test administration. These data were provided during the process to assist the panelists in evaluating their own content expectations and judgments. For example, a panelist's judgment might have been that a particular item was very difficult, but actual student performance data demonstrated that the item was generally easy. One caution that was provided to panelists was that student performance data reflected all levels of achievement. In other words, the empirical item difficulty data was not just representative of the threshold, or *just barely*, students, but on all students sampled.



U.S. History and Government 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
18MC	1	.	.	.	.	9%	.	13%	4%	13%	13%	13%	9%	.	9%	4%	9%	4%	.	.	.	.
13MC	1	.	.	.	.	.	.	13%	4%	13%	13%	13%	4%	.	13%	17%	4%	4%	.	.	.	.
31SCF	1	.	.	.	.	4%	.	.	4%	4%	4%	17%	4%	17%	17%	4%	13%	4%	4%	.	.	.
21MC	1	.	.	.	.	.	4%	.	4%	4%	9%	17%	.	17%	13%	9%	17%	.	.	4%	.	.
27MC	1	.	.	.	.	.	.	4%	9%	4%	9%	13%	13%	22%	4%	9%	4%	.	4%	4%	.	.
36SCF	1	.	.	.	.	.	.	.	.	4%	9%	9%	4%	13%	22%	.	9%	17%	4%	9%	.	.
19MC	1	.	.	.	4%	.	.	9%	17%	13%	9%	13%	13%	13%	.	4%	.	4%	.	.	.	.
20MC	1	.	.	.	.	4%	4%	9%	17%	13%	17%	4%	4%	17%	4%	.	.	4%	.	.	.	.
33SCF	1	.	.	.	.	.	.	.	.	4%	4%	13%	.	17%	17%	9%	4%	22%	.	4%	4%	.
14MC	1	.	.	.	.	.	9%	9%	22%	13%	9%	13%	.	13%	4%	4%	4%	.	.	.	.	.

U.S. History and Government 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
18MC	1	.	.	.	.	.	.	4%	4%	4%	9%	.	4%	22%	4%	17%	4%	4%	17%	.	.	4%
13MC	1	.	.	.	.	.	.	.	4%	.	9%	9%	9%	17%	4%	.	13%	9%	17%	4%	.	4%
20MC	1	.	.	.	.	.	.	4%	4%	.	9%	13%	17%	9%	9%	13%	4%	9%	4%	.	4%	.
14MC	1	.	.	.	.	.	.	.	.	13%	9%	9%	9%	17%	.	13%	13%	4%	9%	4%	.	.
12MC	1	.	.	.	.	.	4%	.	.	.	9%	9%	13%	9%	17%	4%	13%	13%	4%	4%	.	.
19MC	1	.	.	.	.	.	.	4%	.	9%	.	4%	22%	17%	9%	13%	9%	4%	.	4%	4%	.
10MC	1	.	.	.	.	.	4%	.	.	17%	4%	4%	9%	22%	13%	4%	17%	.	.	4%	.	.
4MC	1	.	.	.	.	.	.	.	4%	4%	17%	13%	17%	4%	9%	9%	13%	.	4%	4%	.	.
27MC	1	.	.	.	.	.	.	.	.	4%	.	4%	9%	4%	22%	17%	4%	22%	4%	.	4%	4%
21MC	1	.	.	.	.	.	.	.	.	4%	4%	4%	.	.	13%	17%	22%	13%	17%	.	.	4%

U.S. History and Government 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
18MC	1	.	.	.	.	.	.	.	.	9%	.	.	.	9%	9%	9%	4%	26%	9%	.	22%	4%
10MC	1	.	.	.	.	.	.	.	4%	.	9%	.	4%	9%	4%	9%	22%	13%	13%	13%	.	.
19MC	1	.	.	.	.	.	.	.	.	4%	4%	4%	.	.	.	9%	22%	17%	17%	9%	9%	4%
13MC	1	.	.	.	.	.	.	.	.	4%	.	.	4%	4%	4%	4%	13%	13%	13%	22%	13%	4%
14MC	1	.	.	.	.	.	.	.	.	.	.	9%	4%	9%	4%	9%	9%	22%	17%	.	13%	4%
4MC	1	.	.	.	.	.	.	.	.	.	4%	.	9%	9%	13%	13%	13%	9%	13%	4%	9%	4%
5MC	1	.	.	.	.	.	.	.	.	.	.	17%	.	13%	.	17%	9%	9%	17%	17%	.	.
3MC	1	.	.	.	.	.	.	.	.	4%	.	4%	.	9%	9%	9%	13%	17%	13%	17%	.	4%
20MC	1	.	.	.	.	.	.	.	.	.	9%	.	.	9%	4%	17%	13%	17%	.	22%	9%	.
12MC	1	.	.	.	.	.	.	.	4%	.	.	.	4%	4%	4%	13%	13%	22%	13%	13%	4%	4%

Figure 2. Mock Table of Items with Greatest Variability in Judgments

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 panel-level cut score information and the item judgment variability data (e.g., Figure 2) to discuss. The panelists discussed these data within their tables, but the facilitator led a larger group discussion on items with the 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 7 presents the types of feedback data and at what round they were provided to the panelist.

Table 7. 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 that 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 that what they expected to enter into the standard setting website was what was captured.

    - Committee cut score recommendations and statistics: Panel-level cut score recommendations were the median cut score across all panelists for each performance level. The panelists 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 showed 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 was provided.
- 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, were calculated from the 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, reflected the performance of students who responded to the items during the June 2023 administration.

Level 2 Task Force

Cut scores for Level 2 were recommended by a representative group of four panelists from the larger standard setting panel. This group met at the end of Days 1 and 2 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. The recommended Level 2 cut score and subsequent standard error of judgment were 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. The standard error of judgment (SEJ) was also calculated for the final recommended cut scores to serve as additional information. Table 8 presents a summary of the cut scores for all three rounds, and Table 9 presents the recommended performance level cuts ±3 standard errors of judgment. Figure 3 presents the impact data for all three rounds of standard setting.

Table 8. 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
28	37	47	28.5	38	47	29	38	47

Table 9. Standard Error of Judgment

Performance Level	SEJ	Cut − 3SEJ	Cut − 2SEJ	Cut − 1SEJ	Cut	Cut +1SEJ	Cut +2SEJ	Cut +3SEJ
Level 3	0.42	27.43	27.85	28.28	28.70	29.12	29.55	29.97
Level 4	0.44	36.32	36.76	37.21	37.65	38.09	38.54	38.98
Level 5	0.47	45.60	46.07	46.53	47.00	47.47	47.93	48.40

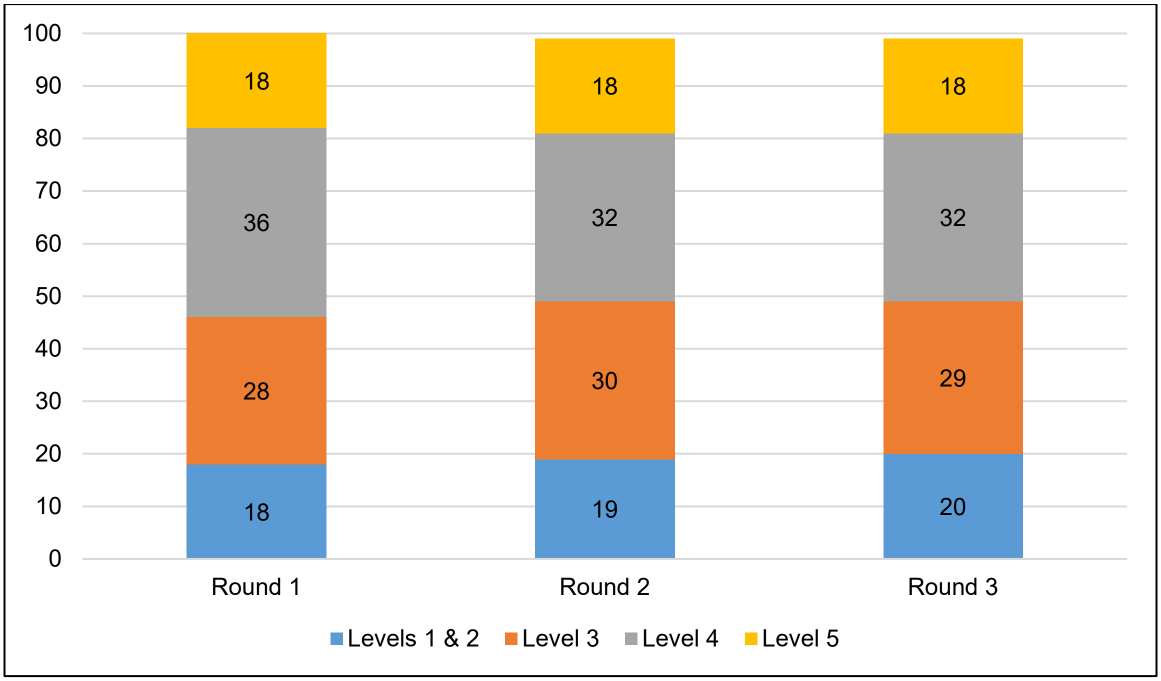


Figure 3. Impact Data by Judgment Round

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 and the materials used and how comfortable they felt about the final recommended cut scores. For the survey questions 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 survey questions used a Likert scale, with different scales of affect (e.g., not confident to very confident, not adequate to very adequate, not useful to very useful) across the evaluation.

One question in particular assessed panelists' confidence in the final cut scores: *How confident do you feel that the final cut score recommendations for U.S. History and Government represent appropriate levels of student performance?* Figure 4 presents the responses of the panelists to this question. As shown in the figure, no panelist indicated that 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 regarding this question exhibits the support from the educators in the process of setting the performance standards for the Regents Examination in United States History and Government (Framework). Some panelists also provided handwritten quotes expressing their gratitude for being a part of this process and desire to participate in other educator panels for the NYS assessments. Appendix E presents all the evaluation results.

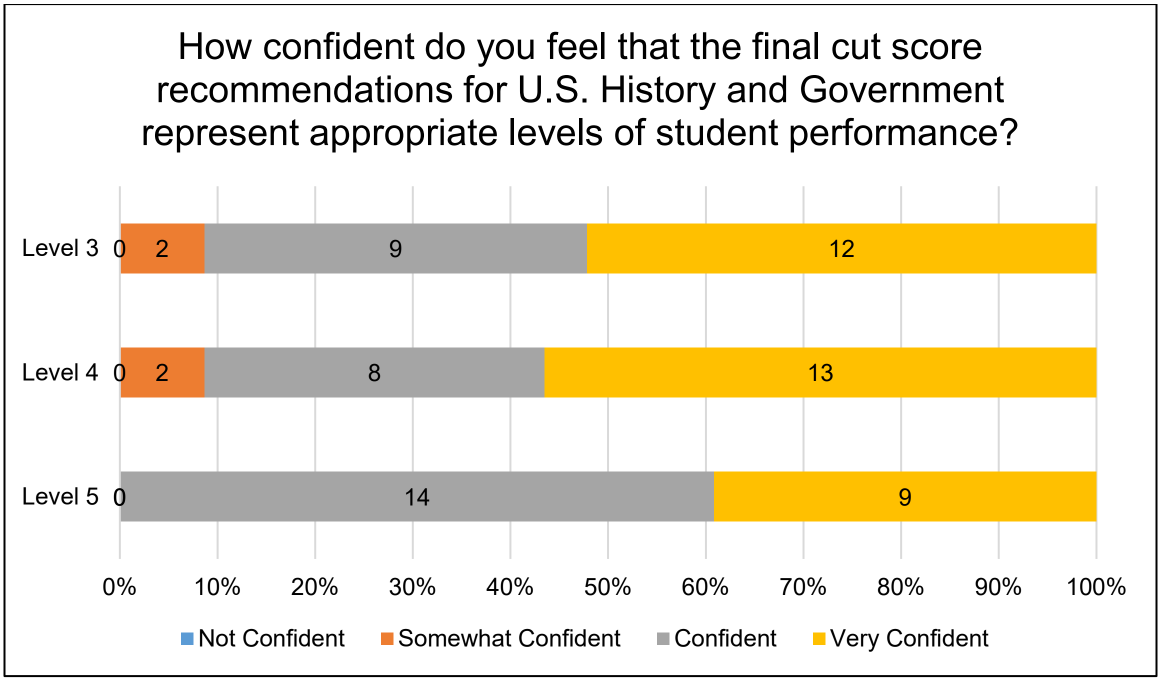


Figure 4. Panelist Confidence in Final Cut Score Recommendations

Final Recommendations

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 method. 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, and the group followed, without incident, instructions delivered by the standard setting facilitator. All activities were formally overseen by the OSA 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 NYSED 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 final approved cut scores were implemented within the scale of measurement used to report student performance on the New York State Regent Examinations. Table 10 presents the approved cuts scores, with subsequent impact data provided in Figure 5.

Table 10. Final Approved Cut Scores

Level 2	Level 3	Level 4	Level 5
23	29	38	47

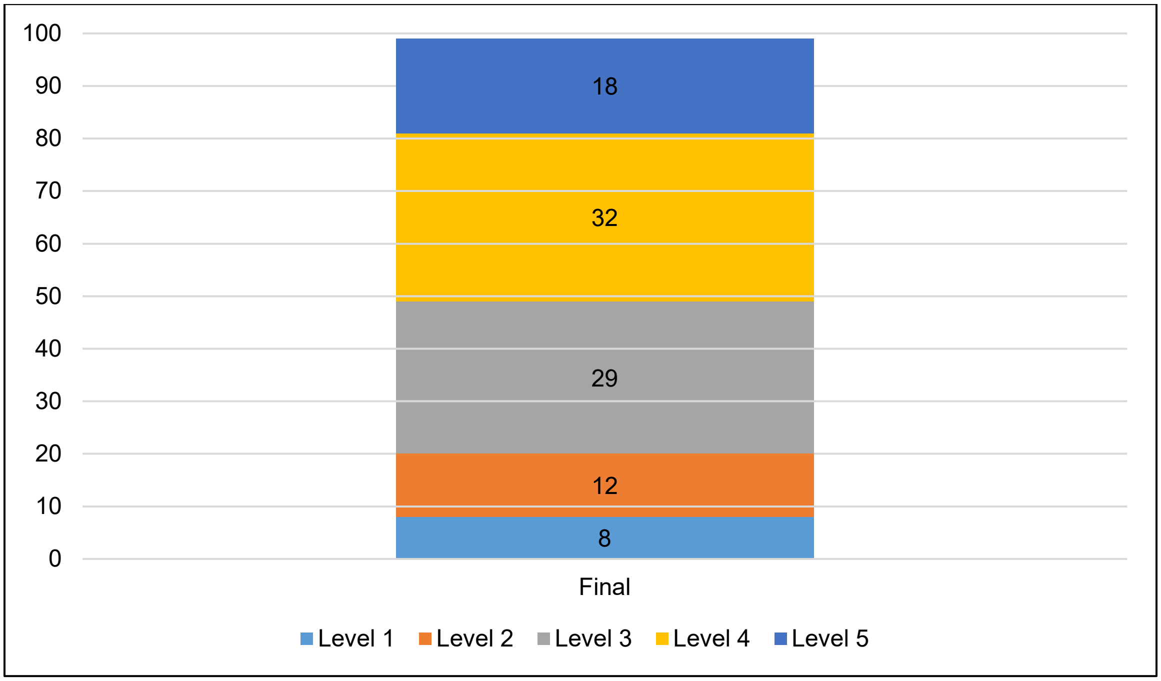


Figure 5. Impact Data based on Final Approved Cut Scores

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). Joint Technical Committee. (2014). Standards for Educational and Psychological Testing. AERA.

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

Egan, K. L., Schneider, M. C., & 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). 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., & 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.

Appendix A: Standard Setting Workshop Agenda

University of the
State of NewYork
State Education
Department

Pearson

Regents Examination in United States History and Government Standard Setting

Agenda

Day 1– June 13, 2023

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 – 7:30 p.m.	Level 2 Task Force

<u>Day 2– June 14, 2023</u>

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
3:30 – 5:00 p.m.	Level 2 Task Force – Cut Score Judgments

Appendix B: Panelist Demographics

While onsite at the standard setting meeting, panelists responded to an information survey to provide demographic and other pertinent information for validity evidence of the standard setting. A total of 24 panelists participated in the standard setting. The survey results have been tabulated below.

Table B.1. Current Position

Responses	#Panelists
Classroom Teacher	23
Administrator (School)	1

Table B.2. Years of Professional Experience in Education

Responses	#Panelists
1–5 years	0
6–10 years	3
11–15 years	3
16–20 years	7
More than 20 years	11

Table B.3. Years of Professional Experience in Teaching U.S. History and Government

Responses	#Panelists
1–5 years	4
6–10 years	7
11–15 years	2
16–20 years	8
More than 20 years	3

Table B.4. Experience with Special Populations

Responses	#Panelists
Students receiving mainstream special education services	24
Students receiving self-contained special education services	3
Students who are English language learners	20
Students who are receiving general education instruction	24
Students who are receiving vocational technical instruction	6

Note. The number of panelists does not add up to 24 because panelists could select all that apply.

Table B.5. Highest Degree Completed

Responses	#Panelists
Master’s degree (M.A., M.S.)	24
Doctoral degree (Ph.D., Ed.D.)	0

Table B.6. Gender

Responses	#Panelists
Male	10
Female	13

Note. Panelists could choose not to answer this question.

Table B.7. Ethnicity

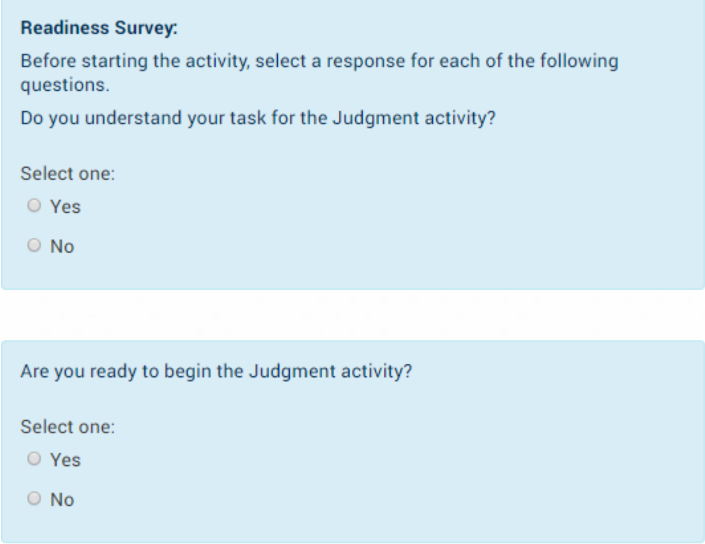
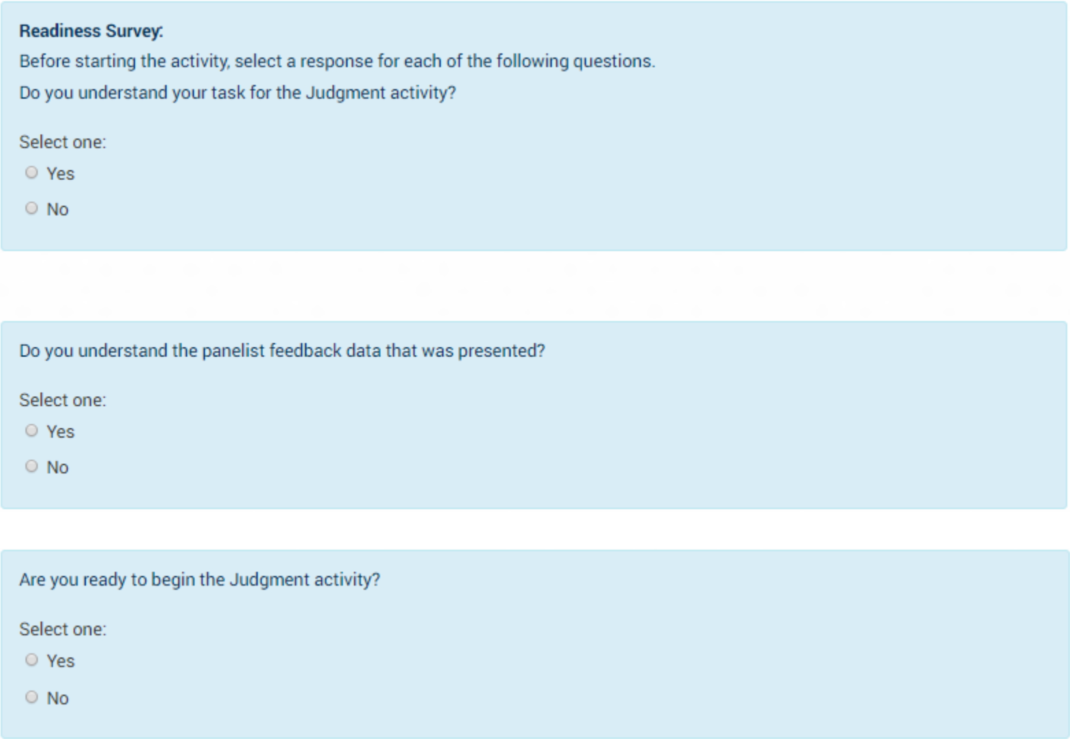
Responses	#Panelists
Hispanic or Latino	1
Not Hispanic or Latino	21

Table B.8. Race

Responses	#Panelists
Asian	1
Black or African American	1
White	19

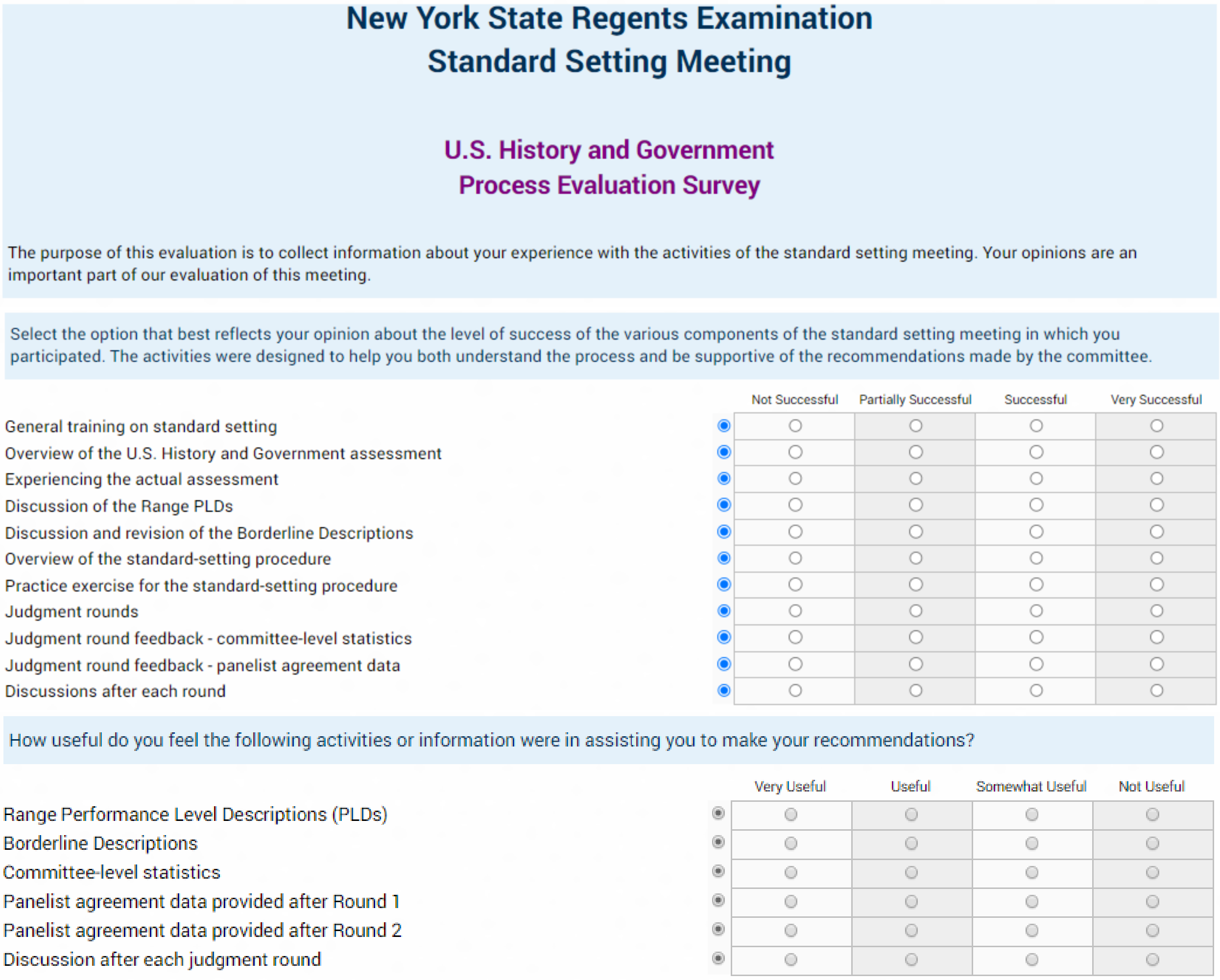
Appendix C: Panelist Readiness Survey

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

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



New York State Regents Examination Standard Setting Meeting

U.S. History and Government Process Evaluation Survey

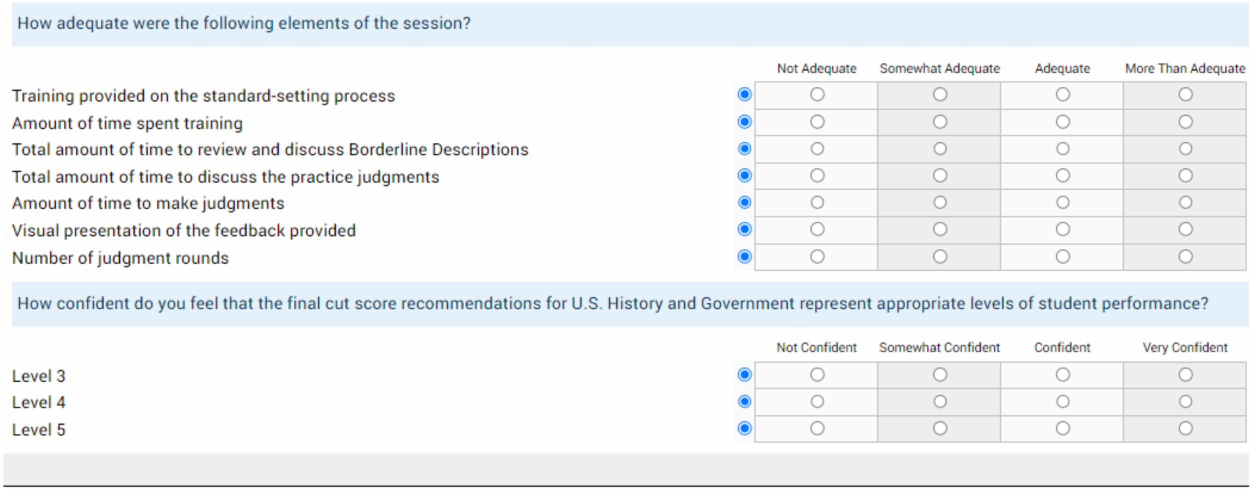
The purpose of this evaluation is to collect information about your experience with the activities of the standard setting meeting. Your opinions are an important part of our evaluation of this meeting.

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.

		Not Successful	Partially Successful	Successful	Very Successful
General training on standard setting	●	○	○	○	○
Overview of the U.S. History and Government 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	●	○	○	○	○

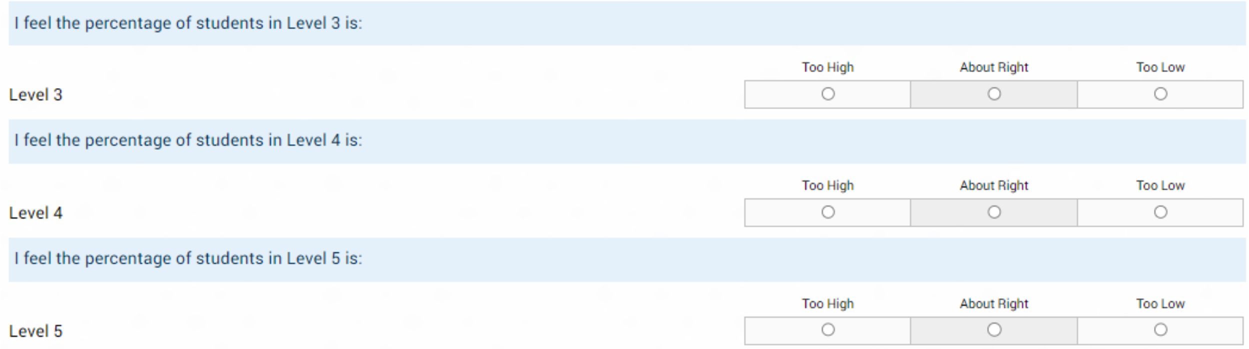


How adequate were the following elements of the session?

		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 U.S. History and Government 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:

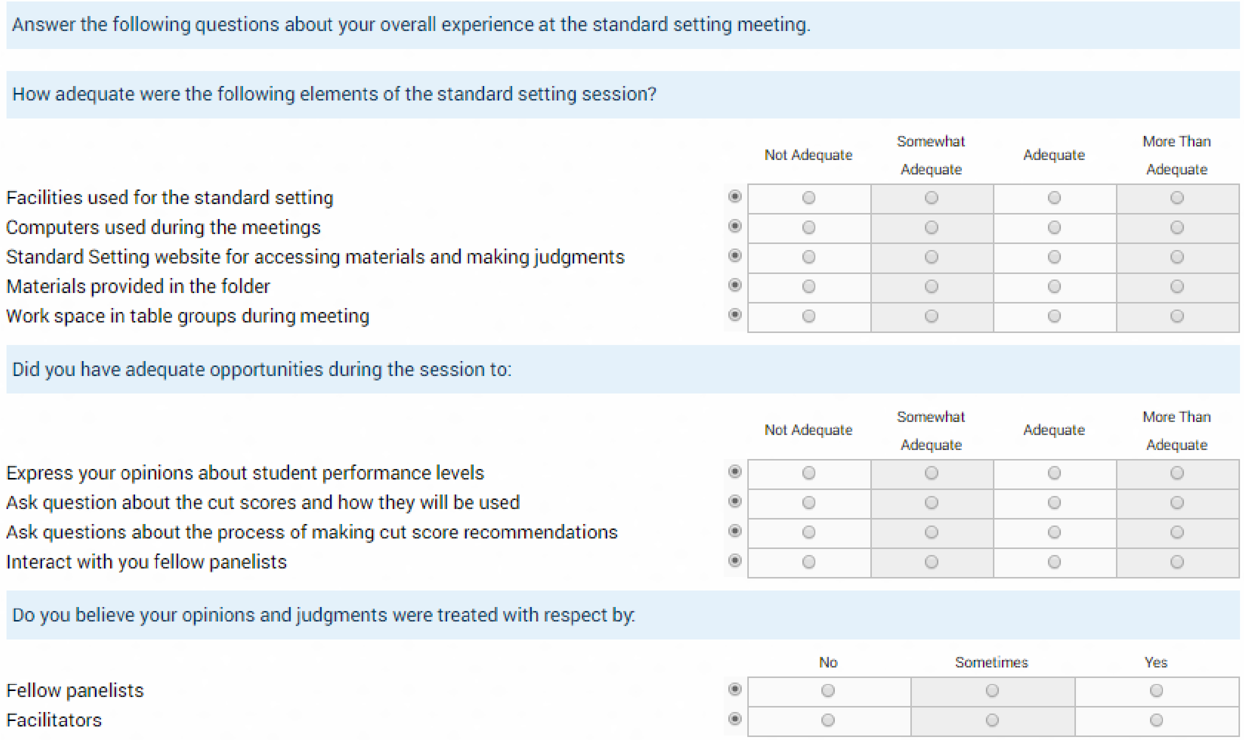
	Too High	About Right	Too Low
Level 3	○	○	○

I feel the percentage of students in Level 4 is:

	Too High	About Right	Too Low
Level 4	○	○	○

I feel the percentage of students in Level 5 is:

	Too High	About Right	Too Low
Level 5	○	○	○



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

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

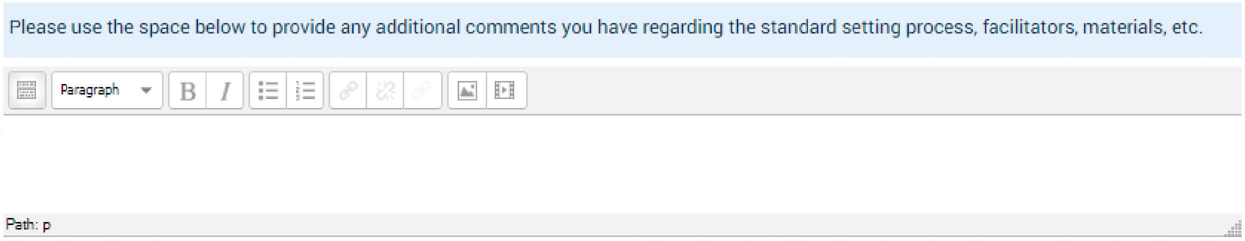
		Not Adequate	Somewhat Adequate	Adequate	More Than Adequate
Facilities used for the standard setting	●	○	○	○	○
Computers used during the meetings	●	○	○	○	○
Standard Setting website for accessing materials and making judgments	●	○	○	○	○
Materials provided in the folder	●	○	○	○	○
Work space in table groups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

Appendix E: Workshop 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	1	8	14
Overview of the United States History and Government examination	0	0	5	18
Experiencing the actual examination	0	0	10	13
Discussion of the Range PLDs	0	3	4	16
Discussion and revision of the Borderline Descriptions	0	1	6	16
Overview of the standard setting procedure	0	2	8	13
Practice exercise for the standard setting procedure	0	0	6	17
Judgment rounds	0	1	2	20
Judgment round feedback - committee-level statistics	0	1	3	19
Judgment round feedback - panelist agreement data	0	1	4	18
Discussions after each round	0	1	3	19

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)	12	8	3	0
Borderline Descriptions	13	6	4	0
Committee-level statistics	17	6	0	0
Panelist agreement data provided after Round 1	18	5	0	0
Panelist agreement data provided after Round 2	18	5	0	0
Discussion after each judgment round	18	4	1	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	1	16	6
Amount of time spent training	0	2	12	9
Total amount of time to review and discuss Borderline Descriptions	0	3	11	9
Total amount of time to discuss the practice judgments	0	0	15	8
Amount of time to make judgments	0	0	8	15
Visual presentation of the feedback provided	0	0	10	13
Number of judgment rounds	0	0	12	11

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

Responses	Not Confident	Somewhat Confident	Confident	Very Confident
Level 3	0	0	14	9
Level 4	0	2	8	13
Level 5	0	2	9	12

Question 5, Question 6, and Question 7: I feel the percentage of students in each Level is:

Responses	Too High	About Right	Too Low
Level 3	0	23	0
Level 4	0	23	0
Level 5	3	19	1

Question 8: 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	11
Computers used during the meetings	0	2	13	8
Standard Setting website for accessing materials and making judgments	0	0	14	9
Materials provided in the folder	0	0	13	10
Work space in table groups during meeting	0	7	9	7

Question 9: 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	0	8	15
Ask questions about the cut scores and how they will be used	0	0	7	16
Ask questions about the process of making cut score recommendations	0	0	6	17
Interact with your fellow panelists	0	0	5	18

Question 10: Do you believe your opinions and judgments were treated with respect by:

Responses	No	Sometimes	Yes
Fellow panelists	0	0	23
Facilitator	0	0	23