Measuring the Power of Learning.™

Research Memorandum
ETS RM–19-06

Providing Threshold Score Recommendations for the Second Phase of Assessments for the *HEIghten*® Outcomes Assessment Suite: A Standard-Setting Study

Wanda D. Swiggett

June 2019

ETS Research Memorandum Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Memorandum series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Memorandum series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

Providing Threshold Score Recommendations for the Second Phase of Assessments for the *HEIghten*® Outcomes Assessment Suite: A Standard-Setting Study

Wanda D. Swiggett
Educational Testing Service, Princeton, New Jersey

June 2019

Corresponding author: W. D. Swiggett, E-mail: WSwiggett@ets.org

Suggested citation: Swiggett, W. D. (2019). *Providing threshold score recommendations for the second phase of assessments for the* HEIghten® *outcomes assessment suite: A standard-setting study* (Research Memorandum No. RM-19-06). Princeton, NJ: Educational Testing Service.

Action Editor: Elizabeth Stone

Reviewers: Richard Tannenbaum and Joseph Rios

Abstract

The *HEIghten*® outcomes assessment suite is made up of 5 assessments designed to measure general student learning outcomes. These modular assessments are designed to provide information to students and institutions, such as student progress on each learning outcome. The standard-setting study for the 3 assessments that make up the first phase of the HEIghten suite (critical thinking, quantitative literacy, and written communication) was conducted in 2016. This report describes the standard-setting study conducted for the 2 remaining HEIghten assessments that make up the second phase of development (civic competency and engagement [CCE] and intercultural competency and diversity [ICD]). The standard-setting panel consisted of 12 college-level educators who teach undergraduate students the transferrable skills measured by these assessments. For 2 forms of each assessment, the panel recommended 2 threshold scores per form, which mark the beginning of the second and third performance categories. Using modified Angoff methods, the panel made 2 rounds of judgments, with feedback and discussion between the rounds. Evaluations were administered after training and at the conclusion of the study. The panelists indicated that their standard-setting training was clear, that the process was easy to follow, that they understood the purpose of the study, and that they supported all of the final recommended scores.

Key words: standard setting, Angoff, *HEIghten*® outcomes assessment suite, civic competency and engagement, intercultural competency and diversity, threshold scores, performance-level descriptors, borderline students, impact data

Acknowledgments

The author would like to thank Kri Burkander, Craig Stief, Dele Kuku, Brenda Richardson, Katrina Roohr, Usama Ali, Douglas Baldwin, Lauren Bauser, Paul Borysewicz, Wyman Brantley, Trina Duke, Joseph Rios, Towanda Sullivan, Edward Wagner, and Jane Wang. Their hard work was instrumental in preparing for and completing this standard-setting study.

Table of Contents

The *HEIghten*® outcomes assessment suite is a collection of modular assessments designed to measure general education student learning outcomes. We designed and conducted the standard-setting study from May 11 to 13, 2016, on two forms of the first three HEIghten assessments—critical thinking, quantitative literacy, and written communication (Swiggett, 2017). A standard-setting study was conducted from December 11 to 13, 2017, on two forms of each of the remaining two assessments: civic competency and engagement (CCE) and intercultural competency and diversity (ICD). The same methodology used during the standard setting for the first phase of tests was used during the second phase.

For the assessments, the HEIghten program considered the recommended scores from the panels as well as other sources of information (such as analyses of data across institutions or student majors) when determining the final threshold scores (Swiggett, 2017). The appropriateness of any adjustment—toward higher or lower scores—can only be evaluated in terms of how well the approved scores support the development and intended use of the test (Geisinger & McCormick, 2010). The terms *student* and *test taker* are used interchangeably throughout this report, which documents the standard-setting process and results for the second phase of assessments.

The HEIghten Outcomes Assessment Suite

The HEIghten Outcomes Assessment suite is a computer-delivered, general education assessment suite of student learning outcomes (Educational Testing Service [ETS], 2018c). These student learning outcomes are skills that students develop throughout their undergraduate years that can be transferred to real-world situations, including future careers. The suite of five assessments is designed to be modular and easy to use. Institutions can select the assessments, or a combination of assessments, in a manner that best meets their needs. Each of the five assessments is designed to have an administration time of 45 minutes, which allows them to be administered during a class period. Other administration options exist, such as administering the assessment online via remote proctoring (ETS, 2018c). The assessments provide institutions with actionable data that can be used for a variety of purposes. For example, institutions can improve teaching and learning by reviewing the evidence of their students' skills, compare the scores to those of other institutions, and use the data to make decisions that impact their educational programs (ETS, 2018c).

Assessments in Phase 2

The two assessments in the Phase 2 development of the outcomes assessment suite—CCE and ICD—launched operationally in January 2018. During the standard-setting study, the panels provided judgments for two forms[1] of each of the Phase 2 assessments, culminating in recommended threshold scores for each assessment. The threshold scores, once accepted by the program, will be used to classify students as developing, proficient, or advanced based on their performance on the test.

Civic competency and engagement. The HEIghten CCE assessment focuses on three key areas: civic competency, civic attitudes, and civic participation. The overall assessment "evaluates college students' knowledge of civic practices and institutions, their skills in understanding and participating in civic life, and their attitudes, preferences, and degrees of engagement in civics more generally" (ETS, 2018a, p. 1). The Likert[2] items in the civic attitudes (30 items) and selected-response[3] items in the civic participation (20 items) dimensions are self-report. They were not included in the standard-setting process because they measure self-reported attitudes, which will vary by person, and do contribute to the reported HEIghten score. Only the 30 selected-response items in the civic competency dimension of the assessment received threshold scores as part of the standard-setting study.

Intercultural competency and diversity. The HEIghten ICD assessment evaluates two central aspects: approach as well as analyze and act. The approach dimension describes "the overall positivity with which an individual views and responds to cross-cultural interactions" (ETS, 2018b, p. 1). The 34 Likert[4] items in the approach dimension were not included in the standard-setting study because they measure individual views and do not contribute to the reported HEIghten score. The analyze and act dimension describes "the ability to take in, evaluate and synthesize relevant information without the bias of preconceived judgments and stereotyped thinking; then, translate that information into action while maintaining control in potentially challenging and stressful situations" (ETS, 2018b, p. 1). The 40 selected-response, situational judgment items in the analyze and act portion of the assessment were part of the standard-setting study.

Performance-Level Descriptors

The performance-level descriptors (PLDs) describe the knowledge, skills, and abilities of students that can be categorized into the developing, proficient, and advanced performance levels

for the assessments.[5] The PLDs are a crucial part of the standard-setting study (Cizek, 2012; Cizek & Bunch, 2007; Perie, 2008). The ETS assessment specialists for each content area developed the PLDs. The PLDs serve as the starting point for the standard-setting panel to develop the borderline-student definitions for proficient and advanced; these definitions operationalize the threshold scores and are used in the judgment-making process. Additionally, text from the PLDs will become part of the score reporting documentation that will be published on the public website of the HEIghten outcomes assessment suite so that test takers and institutions can understand the score reporting. The PLDs, as presented to the standard-setting panel, are listed in Appendix A. The panel provided suggestions to the assessment specialists regarding minor edits (e.g., for clarity). Any future edits to make the final version of the PLDs will retain the essence of the descriptors the panelists viewed.

Standard Setting

The purpose of the standard-setting study for the HEIghten Outcomes Assessment suite was to establish minimum scores that classify test takers into distinct performance levels on specific components of the assessments. The minimum scores are described as threshold scores because they specify the minimum score required to breach the threshold of a performance level. Standard setting is a judgment-based process with no empirically correct passing scores (O'Neill, Buckendahl, Plake, & Taylor, 2007). The concept of how much knowledge or skill must be demonstrated on a test, and embodied by a test score, to reach a level of proficiency or performance is a function of the values and expectations of those involved in setting the standard (O'Neill et al., 2007; Tannenbaum & Katz, 2013). In this value-based context, an evaluation of the credibility and meaningfulness of the passing score—the reasonableness of the passing score—is based on the appropriateness of the standard-setting design and the quality of the implementation of the standard-setting process (Papageorgiou & Tannenbaum, 2016).

Standard setting is part of the collection of validity evidence supporting test development and ultimate use. The design, implementation, and results of a standard-setting study provide evidence supporting the inferences that can be made about the ability of the students who are categorized by the threshold scores. The standard-setting study provides support for the claims that institutions can make about their students (Papageorgiou & Tannenbaum, 2016).

Methodology

Standard setting for the Phase 2 assessments was designed to mirror the methodology implemented for the Phase 1 standard-setting study (Swiggett, 2017). In both phases, the expert panels of judges followed a standard-setting design based on the modified Angoff procedure (Brandon, 2004; Hambleton & Pitoniak, 2006; Plake & Cizek, 2012); panelists made two rounds of judgments with feedback (i.e., data) between rounds (Reckase & Chen, 2012). The feedback prompted discussion from the panelists as they made the second round of judgments. The only difference from the Phase 1 study was that the panelists' judgments were collected electronically, instead of on paper (described later in this section).

As described in *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014), the procedures used during standard setting need to be well documented as part of the validity evidence supporting the recommended scores. The methods used for this study have been used for many standard-setting studies for ETS tests and have been well documented over several years (e.g., Plake & Cizek, 2012; Swiggett, 2017; Tannenbaum, 2011, 2012; Tannenbaum & Kannan, 2015).

In addition to documenting the procedures of the standard-setting methods, the methods were assessed through the use of evaluations of the training and a final evaluation at the conclusion of the study. The panelists' judgments were calculated to determine the panel's recommendation and were also analyzed to provide estimates of the measurement error associated with the judgments. Specifically, the standard deviation of the mean and the standard error of judgment (SEJ; see Appendix B, Technical Note 1) were used as measures of this error. The SEJ is one way of estimating the reliability of a panel's standard-setting judgments. It suggests how likely it would be for several other panels of educators similar in makeup, experience, and training to the current panel to recommend the same scores on the same forms of the assessment (Tannenbaum & Katz, 2013).

Panel of Experts

Standard-setting methods require a panel of experts who have knowledge of the subject matter and experience with the test takers (Cizek & Bunch, 2007). Because it is a process that typically involves discussions among the panel members, it is best to have a panel that is diverse and representative of the field of experts. Given that HEIghten assessments are designed to

measure generic skills among the entire college population, not specific to particular college majors, an interdisciplinary faculty panel was recruited to provide judgments for both assessments. Specifically, faculty with background expertise teaching courses involving diversity, cultural competency, government, and civics were contacted to participate in the study. In addition, because HEIghten is designed to be used at various types of institutions, faculty and administrators from several types of colleges and universities were sought (e.g., public, private, minority serving, 2 or 4 year). The panelists were compensated for their time.

Process and Materials

The panelists were provided with premeeting information about the assessments, the performance levels, and the standard-setting process. At the meeting, panelists familiarized themselves with the assessments and developed two borderline-student definitions for each assessment—proficient and advanced. After training and practice, the panel completed two rounds of judgments, with data presented between the rounds that showed the panel's judgments and impact at each round. The final recommended scores were presented to the panel along with the accompanying impact data, which describe the percentage of students placed into each category based on the panel's recommendations (Margolis & Clauser, 2014). At the conclusion of the study, the panelists were then asked to complete a final evaluation of the standard-setting process. An agenda describing the process followed is included as Appendix C.

Premeeting information. One week before the panelists arrived for the standard-setting study, they were sent descriptions of the HEIghten Outcomes Assessment suite, a link to the HEIghten website, and the *Test-at-a-Glance* documents for each of the two assessments (ETS, 2018a, 2018b). They were also provided with the PLDs for each assessment (see Appendix A). The panelists were asked to consider the PLDs and write notes on the knowledge and skills students at the beginning of the proficient and advanced ranges would be expected to have. It was explained that their notes would be a starting point in their group discussions and that, after they arrived, the entire panel would work together to create full descriptions of these students.

Familiarization with the HEIghten assessments. The panelists reviewed each assessment to gain an understanding of what was being measured, to get a sense of the relative difficulty of the items, and to get a sense of the test takers' experience. They began with an independent review of both forms of the CCE assessment, followed by a group discussion focused on the knowledge and skills the assessment measured and any differences they noticed

between the two forms. The discussion included an evaluation of the learning outcomes college seniors *should* or *must* have when they are ready to enter the workforce. At the conclusion of these discussions, the panelists repeated the test-familiarization process with both forms of the ICD assessment.

Borderline-student definitions. The discussions about the assessments were an important precursor to creating the proficient and advanced borderline-student definitions for each assessment. Although the test takers will be classified as developing, proficient, or advanced, there was no need to create a borderline-student definition for the developing category: A test taker who meets motivation criteria (described subsequently) but who does not earn the minimum threshold score for proficient is categorized as developing.[6] After ETS assessment specialists described the PLDs, the panelists worked in small groups to create definitions. They then worked as a whole group to finalize those definitions by consensus. The completed definitions are included in Appendix D.

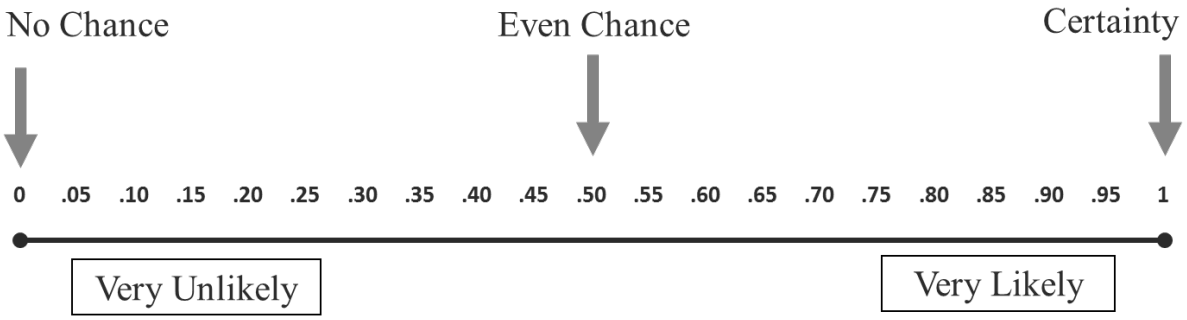

Figure 1. Modified Angoff judgments scale that lists the probabilities the panelists selected as they made their judgments. To aid their decision-making, labels ("No Chance," "Even Chance," "Certainty," "Very Unlikely," and "Very Likely") are added.

Standard-setting judgments. The standard-setting process for the selected-response items was a probability-based modified Angoff method (Brandon, 2004; Hambleton & Pitoniak, 2006; Plake & Cizek, 2012). In this method, each panelist considered each item and then judged the item on the probability that the borderline student would answer the item correctly. The rating scale shown in Figure 1 was used to capture their judgments. The lower the probability value is, the less likely it is that the borderline student would answer the item correctly; the higher the value is, the more likely it is that the borderline student would answer the item

correctly. The panel was told not to consider guessing, fatigue, or anything else unrelated to the knowledge and skills defined in the borderline-student definitions.

Prior to making standard-setting judgments, panelists received training and practice. With each item, panelists reviewed the proficient borderline-student definition (for the CCE assessment) and then the item. They then considered if the borderline proficient student would have a low, moderate, or high chance of answering the question correctly. The facilitator encouraged the panelists to consider the following rules of thumb to guide their decisions:

- Items in the 0–.30 range were those the borderline student would have a low chance of answering correctly.
- Items in the .35–.65 range were those the borderline student would have a moderate chance of answering correctly.
- Items in the .70–1.00 range were those that the borderline student would have a high chance of answering correctly.

Next, panelists decided how to refine their item judgment within the range. For example, if a panelist thought there was a moderate chance that the borderline proficient student would answer the question correctly, the initial decision would be in the .35–.65 range. The second decision for the panelist was to judge if the likelihood of answering it correctly is .35, .40, .45, .50, .55, .60, or .65 and indicate this judgment on his or her rating form. Panelists were reminded not to consider factors unrelated to the student's knowledge and skills. After they completed their judgments for the borderline proficient student, panelists followed the same judgment process for making their judgments for the borderline advanced student on the same item. Using the same scale, the panelists needed to judge the borderline advanced student at a higher probability than their judgments for the borderline proficient student, because the borderline advanced student has a higher level of knowledge and skills.

After the training, panelists made a practice judgment on a subset of CCE items and discussed their judgments and rationales. All panelists completed a posttraining survey to confirm that they had received adequate training and felt prepared to continue. After all panelists confirmed their readiness, they completed their judgments on both forms of the CCE assessment. They did not require further training when they began judgments for the ICD assessment because they were following the same process.

The panelists utilized tablet computers and survey software to complete their judgments for each form of the assessments. Tablets were not used during the Phase 1 standard-setting study because they were not available at that time. Using tablets to collect panelists' judgments had become a regular part of standard-setting procedures by the time of the Phase 2 standard-setting studies. As panelists completed their judgments, minor technical problems (such as being disconnected from the Internet) were quickly resolved and did not cause any delays or errors in the data collection or analysis.

Multiple rounds of judgments. Panelists completed a round of judgments on the first form, followed by the second form, of the CCE assessment. After the first (independent) round of judgments for CCE, the panelists reviewed the item-level feedback from their individual recommendations and from the whole panel. To calculate each panelist's score recommendations for each assessment, their judgments for the selected-response items were summed. The average of the panelists' recommendations was then calculated as the panel's threshold score recommendation. After the panelists completed their second round of judgments for CCE, the process was repeated for both forms of the ICD assessment. The final (Round 2) recommendations were presented to the panel and discussed on the last day of the study.

The panel received feedback on their judgments, including impact data, followed by a discussion. The feedback included the judgments of each panelist from low to high (for each threshold score) and a summary of the panel's judgments, which included the mean, minimum, and maximum judgment values. In addition to the summaries of their judgments, the panel was presented with the impact data based on the field testing of the assessment. Only the seniors' data were included, because the student learning outcomes measured by the assessment are associated with graduating from college. Additionally, students whose performance statistically indicated that they were not motivated to do their best on the assessment were removed from the sample. Students were classified as unmotivated if (a) they failed to complete 75% or more of the test or (b) their response times on at least 20% of the items did not exceed a statistically determined threshold designed to identify rapid guessing (for more details, see Appendix B, Technical Note 2; J. Wang, personal communication, January 10, 2018). The impact data showed the percentage of students who would fall into each of the three categories given the recommended threshold scores from each round of judgments.

The panel discussed their reactions to the judgment data summary and the impact data. Then, for each item and for each threshold, the numbers of judgments in the low, moderate, or high probability ranges were shown. The panel reviewed the item-level data and discussed the rationales for their judgments. The purpose of the discussion was not to encourage panelists to conform to another's judgment but to understand the different relevant perspectives among the panelists. While reviewing the item-level data, panelists made their Round 2 independent judgments. After the second round of judgments, the panel reviewed the post–Round 2 feedback and the impact data. This process was completed for the ICD assessment. Then, they discussed their final recommendations before completing the final evaluation form.

Evaluations of the process. As previously described, the panelists were provided with standard-setting training and practice followed by evaluations. The panelists also completed a final evaluation at the conclusion of the study. These evaluations are used to collect data to evaluate the panelists' judgments of the standard-setting process and final data.

Results

Standard-Setting Panel

Twenty college-level educators from 15 states were nominated from institutions that have had experience with the HEIghten assessments. A total of 12 educators from 11 states agreed to participate in the study. These educators have expertise in teaching the learning outcomes measured in the assessments. The panel represented diverse demographic and professional backgrounds. The panelists taught in small and large colleges and universities, including community colleges. Half of the panelists had been serving as faculty for more than 16 years. The participants taught a large variety of courses (e.g., cross-cultural psychology, intercultural communication, civic literacy, and community-based research) to first-year through senior college students as well as graduate students. Table 1 summarizes the demographics of the standard-setting panel. Names and affiliations of the panelists are listed in Appendix E.

Table 1. Panel Demographics

Demographic	*N*
Current position	
Administrator/department head	6
College faculty	4
Split faculty and administration	2
Race/ethnicity	
White	6
Black or African American	2
Asian or Asian American	3
Prefer not to answer	1
Gender	
Female	11
Male	1
Years of experience (including this year)[a]	
3 or fewer	0
4–7	1
8–11	2
12–15	2
16 or more	6
Students taught[b]	
Freshmen/first-years	5
Sophomores	7
Juniors	9
Seniors	10
Other (including graduate students)	4

[a]An administrator/department head did not indicate her years of experience. [b]Panelists were asked to select all that apply.

Standard-Setting Judgments

Table 2 summarizes the final standard-setting judgments (Round 2) of the panel. Panelist-level results for Rounds 1 and 2, including standard deviation and SEJ of the panel's judgments, are presented in Appendix F (Tables F1–F4), along with impact data based on their judgments from each round (Figures F1–F4). The estimates of measurement error associated with the judgments support the consistency of the panel's judgments (Papageorgiou & Tannenbaum, 2016; Tannenbaum & Kannan, 2015; Tannenbaum & Katz, 2013).

Round 1 judgments are made without discussion among the panelists. The most variability in judgments, therefore, is typically present in the first round. Round 2 judgments, however, are informed by panel discussion; thus it is common to see a decrease in both standard deviation and SEJ. This decrease—indicating convergence among the panelists' judgments—was observed for most of the threshold scores for the forms of each assessment. For both forms of ICD, however, the SEJ increased for the advanced judgments from Rounds 1 to 2. The removal of a low outlier (described later in this section) for Form 2 would show a decrease in the SEJ

from Rounds 1 to 2. The judgments from the same panelist were also the lowest for Form 1 but did not meet the criteria for classification as an outlier. Increases in SEJ after Round 1 are not typical and indicate that the panelists did not come to a greater consensus after discussion. The panel's recommended threshold scores are the Round 2 mean scores.

Table 2. Recommended Threshold Scores and Standard Errors of Judgment by Round

Assessment and form number	Proficient Round 1	Proficient Round 2	Advanced Round 1	Advanced Round 2
Civic Competency and Engagement 1	12.18 (0.88)	12.20 (0.74)	23.53 (0.66)	23.65 (0.61)
Civic Competency and Engagement 2	11.13 (0.97)	11.38 (0.81)	22.59 (0.70)	22.76 (0.56)
Intercultural Competency and Diversity 1	16.39 (0.83)	16.75 (0.75)	31.95 (0.38)	32.21 (0.44)
Intercultural Competency and Diversity 2	15.36 (0.97)	15.66 (0.87)	31.80 (0.39)	31.95 (0.49)

Note. All threshold score recommendations are listed as raw scores. The maximum raw scores that can be earned are 30 for civic competency and engagement and 40 for intercultural competency and diversity. Standard errors of judgment are in parentheses.

Table 3 presents the estimated conditional standard errors of measurement (CSEM) around the recommended passing scores (see Appendix B, Technical Note 3). The standard error of measurement represents the uncertainty associated with the test score. The CSEM provided are estimates.

Table 3. Final Recommended Scores and Conditional Standard Errors of Measurement

Assessment and form code	Proficient	Advanced
Civic Competency and Engagement 1	13 (2.76)	24 (2.23)
Civic Competency and Engagement 2	12 (2.73)	23 (2.36)
Intercultural Competency and Diversity 1	17 (3.17)	33 (2.43)
Intercultural Competency and Diversity 2	16 (3.14)	32 (2.56)

Note. All threshold score recommendations are listed as raw scores based on the panel's final (Round 2) judgment. The scores are rounded to the next highest number. (A score of 11.2, for example, means that a test taker must correctly answer *more than* 11 selected-response items, which is 12 items.) The maximum raw scores that can be earned are 30 for civic competency and engagement and 40 for intercultural competency and diversity. Conditional standard errors of measurement are in parentheses.

Outlier analyses were completed on the panel's Round 2 judgments (see Appendix B, Technical Note 4). A low outlier was found for the advanced level of CCE, Form 2. A low outlier was also found for the advanced level of ICD, Form 2. By removing the outliers, the rounded study values would increase by 1 point, and the SEJ for each test would decrease. The study values and CSEM would match the recommended values from Form 1 of each test. These data are listed in Table 4; additional information is included in Appendix F.

Table 4. Final Recommended Scores, Conditional Standard Errors of Measurement, and Standard Errors of Judgment With Outliers Removed

Assessment and form code	Advanced	CSEM	SEJ
Civic Competency and Engagement 2	24	2.33	0.48
Intercultural Competency and Diversity 2	33	2.43	0.37

Note. CSEM = conditional standard error of measurement; SEJ = standard error of judgment.

Evaluations

The panelists were given an evaluation after they were provided with training, practice, and discussion. All of the panelists verified that they understood the process and confirmed their readiness to proceed. Final evaluations were administered at the conclusion of the standard-setting study. The results are shown in Appendix G. Table G1 provides overall results. All panelists strongly agreed or agreed that they understood the purpose of the study. All of the panelists strongly agreed or agreed that the facilitator's instructions and explanations were clear. Table G2 summarizes factors that influenced panelists' decisions. All of the panelists reported that the descriptions of the borderline students were at least somewhat influential in guiding their standard-setting judgments. The panel also reported that the between-round discussions were at least somewhat influential in guiding their judgments as was their own professional experience.

After the panelists were provided the final recommended threshold scores, they were asked to complete the questions on the final evaluations pertaining to those scores. They were asked (a) if they believed that the final recommendations were too high, too low, or about right and (b) if they supported the final recommendations of the panel. One panelist needed to leave early and did not see the final recommendations. As a result, that panelist could not answer those questions. Most of the remaining 11 panelists indicated that the final recommended scores were about right for each level of each assessment, with some stating that the recommendations were too high (one panelist for the CCE advanced level) or too low (two panelists for the CCE proficient level and three for the ICD proficient level). For the advanced level of the ICD assessment, only six indicated that the recommended scores were about right. Three selected too low and two selected too high on the evaluation (see Appendix G, Table G3).

The 11 panelists who completed the final recommendations, however, indicated that they supported the panel's final recommendations for both levels of the CCE assessments (see Appendix G, Table G4). For the ICD assessment, 10 of the panelists supported the final recommendations of the panel. One panelist did not support the final recommendation for ICD

proficient level, and one panelist did not respond to the question regarding the ICD advanced level. The latter panelist verbally expressed uncertainty about supporting the panel's recommendation.

Discussion

We designed and conducted the standard-setting study described in this report to support the decision-making process for the HEIghten program in establishing threshold scores for the Phase 2 assessments. Standard setting is a judgment-based process that relies on the considered judgments of subject-matter experts. The confidence placed on the recommended score is bolstered by procedural evidence—the quality of the standard-setting study—and internal evidence: the likelihood of replicating the recommended threshold scores (Kane, 1994, 2001).

The makeup of the panel is an essential part of the standard-setting study. This panel was made up of 12 college-level educators with diverse backgrounds and experiences, with experience with the content measured on the assessments and also with students who would take the assessments. The size of the panel for this study is considered acceptable (e.g., Plake, Impara, & Irwin, 2000; Tannenbaum & Kannan, 2015), and panelists were able to follow the process and understand the training, as documented in the evaluations. Additionally, they provided judgments that were consistent with each other and across forms of the assessments.

Procedural evidence often comes from panelists' responses to the training and end-of-study evaluations (Cizek, 2012; Cizek & Bunch, 2007; Papageorgiou & Tannenbaum, 2016). The panelists completed an evaluation after their training and also at the conclusion of their standard-setting study. After training, they provided feedback regarding the quality of their training and readiness to make their judgments. The final evaluation asked the panelists to provide feedback about the quality of the standard-setting implementation and the factors that influenced their decisions. The responses to the evaluations provided evidence of the validity of the standard-setting process and, as a result, evidence of the reasonableness of the recommended threshold scores. It is typical to see a small number of panelists indicate disagreement with the panel's final recommendations. Considering that the recommendations are based on the judgments of the entire panel, it is also typical for panelists who disagree with the final score to support the recommendations nonetheless.

Internal evidence (consistency) addresses the likelihood of replicating the recommended threshold scores. For single-panel standard-setting studies, the standard error associated with the

recommended scores can approximate the replicability of the results (Cizek & Bunch, 2007; Kaftandjieva, 2010; Tannenbaum & Kannan, 2015). This SEJ is an index of the extent to which the threshold scores would vary if the study were repeated with different panels (Tannenbaum & Katz, 2013). The smaller the value is, the less likely it is that other panels would recommend a significantly different threshold score. The CSEM of the forms of each assessment also supported the consistency of the scores associated with those forms. The assessments were developed to be of equivalent difficulty, and the independent judgments from the panelists provided evidence that this goal was reached.

The HEIghten program requested that a standard-setting study be conducted on the assessments so that they could have independent, expert evidence supporting the inferences that can be made about the scores and the performance levels. In addition to now having threshold scores for the two Phase 2 HEIghten assessments, the PLDs that were developed will also be used on the score reports. They provide a rich description of the performance levels for the college students and institutions interested in the assessments and allow stakeholders to understand the meaning behind the test takers' score categorizations. The HEIghten program reviewed the recommendations of the standard-setting panel. At the time of the writing of this report, final decisions regarding the scores had not yet been made.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Brandon, P. R. (2004). Conclusions about frequently studied modified Angoff standard-setting topics. *Applied Measurement in Education*, *17*, 59–88. https://doi.org/10.1207/s15324818ame1701_4

Cizek, G. J. (Ed.). (2012). *Setting performance standards: Foundations, methods, and innovations*. New York, NY: Routledge. https://doi.org/10.4324/9780203848203

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage. https://doi.org/10.4135/9781412985918

Educational Testing Service. (2018a). *HEIghten® Civic Competency and Engagement: Test-at-a glance*. Retrieved from https://www.ets.org/s/heighten/pdf/heighten_cce_test_at_a_glance.pdf

Educational Testing Service. (2018b). *HEIghten® Intercultural Competency and Diversity: Test-at-a glance*. Retrieved from https://www.ets.org/s/heighten/pdf/heighten_icd_test_at_a_glance.pdf

Educational Testing Service. (2018c). *Introducing the HEIghten® Outcomes Assessment suite*. Retrieved from https://www.ets.org/heighten

Geisinger, K. F., & McCormick, C. M. (2010). Adopting cut scores: Post-standard-setting panel considerations for decision makers. *Educational Measurement: Issues and Practice*, *29*(1), 38–44. https://doi.org/10.1111/j.1745-3992.2009.00168.x

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433–470). Westport, CT: American Council on Education/Praeger.

High, R. (2000). Dealing with "outliers": How to maintain your data's integrity. *Computing News*, *15*(3), 14–16.

Kaftandjieva, F. (2010). *Methods for setting cut scores in criterion-referenced achievement tests: A comparative analysis of six recent methods with an application to tests of reading in EFL*. Arnhem, Netherlands: CITO.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research*, *64*, 425–462. https://doi.org/10.3102%2F00346543064003425

Kane, M. T. (2001). So much remains the same: Conceptions and status of validation in setting standards. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 53–88). Mahwah, NJ: Erlbaum.

Lord, F. M. (1984). Standard errors of measurement at different ability levels. *Journal of Educational Measurement*, *21*, 239–275. https://doi.org/10.1111/j.1745-3984.1984.tb01031.x

Margolis, M. J., & Clauser, B. E. (2014). The impact of examinee performance information on judges' cut scores in modified Angoff standard-setting exercises. *Educational Measurement: Issues and Practice*, *33*(1), 15–22. https://doi.org/10.1111/emip.12025

O'Neill, T. R., Buckendahl, C. W., Plake, B. S., & Taylor, L. (2007). Recommending a nursing-specific passing standard for the IELTS examination. *Language Assessment Quarterly*, *4*, 295–317. https://doi.org/10.1080/15434300701533562

Papageorgiou, S., & Tannenbaum, R. J. (2016). Situating standard setting within argument-based validity. *Language Assessment Quarterly*, *13*, 109–123. https://doi.org/10.1080/15434303.2016.1149857

Perie, M. (2008). A guide to understanding and developing performance-level descriptors. *Educational Measurement: Issues and Practice*, *27*, 15–29. https://doi.org/10.1111/j.1745-3992.2008.00135.x

Plake, B. S., & Cizek, G. J. (2012). Variations on a theme: The modified Angoff, extended Angoff, and yes/no standard setting methods. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 181–199). New York, NY: Routledge.

Plake, B. S., Impara, J. C., & Irwin, P. M. (2000). Consistency of Angoff-based predictions of item performance: Evidence of technical quality of results from the Angoff standard setting method. *Journal of Educational Measurement*, *37*, 347–355. https://doi.org/10.1111/j.1745-3984.2000.tb01091.x

Reckase, M. D., & Chen, J. (2012). The role, format, and impact of feedback to standard setting panelists. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 149–164). New York, NY: Routledge.

Swiggett, W. D. (2017). *Providing threshold score recommendations for the first three tests of the HEIghten Outcomes Assessment suite: A standard-setting study* (Research Memorandum No. RM-17-06). Princeton, NJ: Educational Testing Service.

Tannenbaum, R. J. (2011). Setting standards on *The Praxis Series*™ tests: A multistate approach. *R&D Connections*, *17*, 1–9.

Tannenbaum, R. J. (2012). A multistate approach to setting standards: An application to teacher licensure tests. *CLEAR Exam Review*, *23*(1), 18–24.

Tannenbaum, R. J., & Kannan, P. (2015). Consistency of Angoff-based standard-setting judgments: Are item judgments and passing scores replicable across different panels of experts? *Educational Assessment*, *20*, 66–78. https://doi.org/10.1080/10627197.2015.997619

Tannenbaum, R. J., & Katz, I. R. (2013). Standard setting. In K. F. Geisinger (Ed.), *APA handbook of testing and assessment in psychology: Vol. 3. Testing and assessment in school psychology and education* (pp. 455–477). Washington, DC: American Psychological Association.

Wise, S. L., & Kong, X. (2005) Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18,* 163–183. https://doi.org/10.1207/s15324818ame1802_2

List of Appendices

Appendix A. Performance-Level Descriptors

HEIghten Civic Competency and Engagement: Civic Competency[7] Performance-Level Descriptions

Advanced

A typical student at the advanced level has demonstrated the ability to

- understand how knowledge of fundamental government institutions and processes enables more effective civic participation
- evaluate strengths and weaknesses of potential solutions to public debates over major political or social issues
- apply appropriate evaluative standards to different types of media (e.g., social media, online newspapers, network television) used to convey political information
- relate historical events and trends to current political discourse and debates
- understand the potential effects of laws or policies on different communities or groups
- apply principles from foundational political documents to current political or community issues
- understand fundamental principles of democratic processes, civil rights, and the rule of law
- understand legal aspects of citizenship, voting, and representation
- consistently distinguish evidence-backed facts from unsubstantiated opinions
- apply appropriate ethical and/or democratic principles to political decisions or practices
- consistently choose the most appropriate mode of participation to resolve political or community issues
- analyze social or political systems to plan processes of problem solving and public action
- relate national policy and political events to a global or international perspective

Proficient

A typical student at the proficient level has demonstrated the ability to

- understand fundamental government institutions and processes
- identify potential solutions to public debates over major political or social issues
- recognize differences in how political information is conveyed through different types of media (e.g., social media, online newspapers, network television)
- understand current political discourse and debates
- understand the potential effects of laws or policies
- understand the relationship of foundational political documents to current political or community issues
- understand fundamental principles of democratic processes, civil rights, and the rule of law
- understand legal aspects of citizenship, voting, and representation
- consistently distinguish evidence-backed facts from unsubstantiated opinions
- apply appropriate ethical and/or democratic principles to political decisions or practices
- consistently choose the most appropriate mode of participation to resolve political or community issues

Developing

A typical student at the developing level may sometimes

- display knowledge of fundamental government institutions and processes
- identify different positions in public debates over major political or social issues
- recognize when political information and opinion are being conveyed through different types of media (e.g., social media, online newspapers, network television)
- display some knowledge of current political discourse and debates
- recognize the effects of laws or policies

- comprehend the content of foundational political documents
- display knowledge of fundamental principles of democratic processes and civil rights
- display knowledge of the legal aspects of citizenship, voting, and representation
- distinguish evidence-backed facts from unsubstantiated opinions
- identify ethical and/or democratic principles relevant to political decisions or practices
- choose an appropriate mode of participation to resolve political or community issues

HEIghten Intercultural Competency and Diversity Analyze and Act[8] Performance-Level Descriptions

Note on Interpreting Scores

Scores on the Analyze and Act dimension reflect the reactions of test takers to descriptions of interactions among culturally different others. Scores may not reflect how individuals will actually respond or perform in real-world situations.

Analyze and Act (Performance-Level Descriptors)

Advanced

In responding to descriptions of interactions with culturally different others, test takers at this level are highly aware of/able to identify

- the impact of their own culture, values, preferences, and previous experiences on their cognitive, emotional, and behavioral responses
- how certain behaviors or actions may be interpreted by other people
- how nonverbal behaviors or cues may signal certain feelings, thoughts, or intentions
- others' responses to their own actions and signals
- others' physical, verbal, and nonverbal behaviors and cues during a social interaction
- others' potential viewpoints

- how preconceived judgments and stereotyped thinking can interfere with information processing
- how to use declarative cultural knowledge to enhance interactions (with culturally different others)
- the importance of monitoring and revising personal behavior to engage in culturally appropriate behavior and to avoid culturally inappropriate behavior
- the importance of monitoring and revising emotions in an automatic or controlled manner

Proficient

In responding to descriptions of interactions with culturally different others, test takers at this level are moderately aware of/able to identify

- the impact of their own culture, values, preferences, and previous experiences on their cognitive, emotional, and behavioral responses
- how certain behaviors or actions may be interpreted by other people
- how nonverbal behaviors or cues may signal certain feelings, thoughts, or intentions
- others' responses to their own actions and signals
- others' physical, verbal, and nonverbal behaviors and cues during a social interaction
- others' potential viewpoints
- how preconceived judgments and stereotyped thinking can interfere with information processing
- how to use declarative cultural knowledge to enhance interactions (with culturally different others)
- the importance of monitoring and revising personal behavior to engage in culturally appropriate behavior and to avoid culturally inappropriate behavior
- the importance of monitoring and revising emotions in an automatic or controlled manner

Developing

In responding to descriptions of interactions with culturally different others, test takers at this level are not very aware of/able to identify

- the impact of their own culture, values, preferences, and previous experiences on their cognitive, emotional, and behavioral responses
- how certain behaviors or actions may be interpreted by other people
- how nonverbal behaviors or cues may signal certain feelings, thoughts, or intentions
- others' responses to their own actions and signals
- others' physical, verbal, and nonverbal behaviors and cues during a social interaction
- others' potential viewpoints
- how preconceived judgments and stereotyped thinking can interfere with information processing
- how to use declarative cultural knowledge to enhance interactions (with culturally different others)
- the importance of monitoring and revising personal behavior to engage in culturally appropriate behavior and to avoid culturally inappropriate behavior
- the importance of monitoring and revising emotions in an automatic or controlled manner

Appendix B. Technical Notes

1. Standard Error of Judgment

The SEJ is one way of estimating the reliability or consistency of a panel's standard-setting judgments. It indicates how likely it would be for several other panels of educators similar in makeup, experience, and standard-setting training to the current panel to recommend the same threshold score on the same form of the assessment. A SEJ assumes that panelists are randomly selected and that standard-setting judgments are independent. It is seldom the case that panelists are randomly sampled, and only the first round of judgments may be considered independent. The SEJ, therefore, likely underestimates the uncertainty of threshold scores (Tannenbaum & Katz, 2013).

The SEJ is the standard deviation (*SD*) of the panelists' judgments divided by the square root of the number of panelists (*n*):

$$\text{SEJ} = \frac{\text{SD}}{\sqrt{n}} \ .$$

2. Sample Rules for Motivated Filtering

The data analysts deleted from each sample those who met either of the following rules.

Rule 1 (75% Rule)

Any test taker who did not complete at least 75% of the items within the test section under analysis was eliminated from that analysis. Note that for the Phase 2 tests, the completion percentage criterion is applied to each section separately. For example, each ICD test form has two sections: 36 Likert-scale items and 40 dichotomously scored items (i.e., all of the selected-response items).

Rule 2 (Motivation Rule; for Dichotomously Scored Items)

The data analyst applied a series of checks on the students' responses, paying particular attention to the latency, which is the length of time required to produce the response. The theory is that responses with very short latencies are more often the result of random guessing rather than valid test-taking behavior. The result of these checks is to classify each student's score record as either motivated (coded as 1) or unmotivated (coded as 0). To implement this

classification rule, we required two thresholds, one imposed at the item level (rapid-guessing threshold) and the second at the form level (response time effort threshold; Wise & Kong, 2005). The process works as follows:

1. Classify each response a student provides as valid or invalid by comparing the latency for item *i* and test taker $j(l_{ij})$ to the item threshold t_i, which is defined as

$$t_i = .10 \times \bar{t}_i,$$

where $\bar{t}_i$ is the sample median item response time for item *i*. The solution behavior for item *i* and test taker $j(SB_{ij})$ is computed as

$$SB_{ij} = \begin{cases} 1 & if\ l_{ij} \geq t_i \\ 0 & \text{otherwise.} \end{cases}$$

This process holds for nonmissing responses. For missing responses, $SB_{ij} = 0$. Conceptually, the response time threshold t_i for a given item defines the boundary between solution behavior and rapid guessing. Response times under that threshold are taken to indicate that a test taker would not have had a reasonable chance to read the item and identify the correct answer.

2. Compute the response time effort index for test taker $j(\text{RTE}_j)$, which is defined as

$$\text{RTE}_j = \sum_{i=1}^{I} (SB_{ij} / I),$$

where *I* is equal to the number of responses provided by the test taker.

3. Classify the motivation level of test taker $j(\text{MOT}_j)$ by comparing RTE_j to the RTE threshold (RTE_T), which is equal to .80 for all forms. Specifically, if $\text{RTE}_j > \text{RTE}_T$, (MOT_j) is equal to 1; otherwise, it is equal to 0. Students with a motivation level of 0 were removed from the reporting.

3. Estimated Conditional Standard Error of Measurement

The estimated CSEM for a test consisting of dichotomously scored items is computed from the study value (SV) of the recommended passing score and the number of dichotomously scored items (n) on the test (see Lord, 1984):

$$\text{CSEM} = \sqrt{\left[\text{SV}\left(n - \text{SV}\right)\right] \,/\, \left(n - 1\right)}\,.$$

4. Outlier Analysis

An outlier analysis of the Round 2 data points was reviewed for each form of both tests and at each performance level. Judgments that were above or below 1.5 times the interquartile range for the round were identified as outliers (High, 2000). Removal of the outliers is at the discretion of the HEIghten program. All panelists were observed to be following standard-setting methods consistent with the training.

Appendix C. HEIghten Outcomes Assessment Suite Standard-Setting Agenda

December 11–13, 2017

Day 1

8:00 A.M.	Welcome and introductions
	Overview of standard setting
	Overview of the tests
	Discussion about student outcomes that are essential for success
	Break
	Review Form A of CCE test and self-score
	Review Form B of CCE test and self-score
	Discuss the content measured on the forms
	Lunch
	Review Form A of ICD test and self-score
	Review Form B of ICD test and self-score
	Break
	Discuss the content measured on the forms
5:00 P.M.	Collect materials; end of Day 1

Day 2

8:00 A.M.	Overview for Day 2
	Overview of performance-level descriptions
	Define borderline students (BLS) in relation to what is measured on Civic Competency and Engagement (CCE) test * Define borderline proficient student * Define borderline advanced student
	Break

	Define BLS in relation to what is measured on Intercultural Competency and Diversity (ICD) test * Define borderline proficient student * Define borderline advanced student
	Lunch
	Round 1 standard-setting training and practice for selected-response judgments
	Complete Round 1 judgments for CCE tests (proficient and advanced BLS)
	Break
	Data presentation and discussions
	Complete Round 2 judgments for CCE tests (proficient and advanced BLS)
5:00 P.M.	Collect materials; end of Day 2

Day 3

8:00 A.M.	Overview for Day 3
	Complete Round 1 judgments for ICD tests (proficient and advanced BLS)
	Data presentation and discussions for ICD tests
	Complete Round 2 judgments for ICD tests (proficient and advanced BLS)
	Break
	Final results and discussion
	Final evaluations
	Collect materials; end of Day 2
1:00 P.M.	Lunch

Appendix D. Borderline-Student Definitions

Civic Competency and Engagement Assessment

Civic Competency Dimension

The borderline proficient student

- can describe the roles of fundamental government institutions and processes
- can articulate the general role of the Constitution in U.S. government
- can differentiate between evidence-backed facts and unsubstantiated opinions and can identify where one is likely to find factual information
- can explain in basic terms the general and fundamental principles of democratic participation (e.g., voting systems, political discourse, civic participation), civil rights, and rule of law
- can accurately judge clear cases of unethical versus ethical action in the public sphere
- recognizes that potential solutions can exist to major political or social issues

The borderline advanced student

- can identify connections between different levels of government institutions to engage in appropriate civic and democratic participation (e.g., voting, engaging in political discourse)
- recognizes connections between foundational documents, historical events, and trends to current political discourses and debates
- recognizes how U.S. policy/political events may have global effects, and the reverse
- can apply facts-based information to problem solving and public action
- can identify strengths and weaknesses of potential solutions to major political or social issues
- can examine ambiguous cases of unethical versus ethical action in the public sphere at a basic level
- can identify strengths and limitations of different approaches to government

Intercultural Competency and Diversity Assessment

Analyze and Act Dimension

The borderline proficient student

- recognizes that different cultures may interpret behaviors differently
- can identify that one's culture, values, preferences, stereotypes, and experience influence how one acts
- has obtained basic, foundational, and factual knowledge that can be used to interact in intercultural situations and make comparative judgments
- recognizes the need to reserve judgment and be open-minded when faced with cultural differences
- recognizes that there are culturally appropriate and inappropriate behaviors that could require self-monitoring

The borderline advanced student

- can identify specific actions that different cultures may interpret differently
- can look at others in social situations and detect key cues in physical, verbal, and nonverbal behavior
- recognizes preconceived judgments and stereotypes and how they influence thinking and begins to change how he or she thinks and acts
- can use awareness that one's culture, values, preferences, and experiences influence how one thinks, feels, and acts to enhance interactions
- can be open to changing behavior based on information presented
- can monitor his or her emotions (to refrain from jumping to conclusions, to observe and collect social information)

Appendix E. Participating Panelists and Affiliation

Panelist	Affiliation and state
Carolyn Barber	University of Missouri, Kansas City (MO)
Sonia Dalmia	Seidman College of Business, Grand Valley State University (MI)
William K. Gabrenya Jr.	Florida Institute of Technology (FL)
Jennifer Johnson Kebea	Drexel University (PA)
Molly Kerby	Western Kentucky University (KY)
Annette S. Kluck	Auburn University (AL)
Andrea Lynch	Mercer County Community College (NJ)
Winnie Mucherah	Ball State University (IN)
Mitsu Narui	Capital University (OH)
Cheryl Norman	University of Northwestern–St. Paul (MN)
Traci O'Brien	Auburn University (AL)
Yifang Zhang	University of Oregon (OR)

Note. Panelists were randomly assigned ID numbers to protect the anonymity of their data. Panelists provided permission for their names and affiliations to be listed in this report.

Appendix F. Threshold Scores and Impact Data per Round of Judgments

Table F1. HEIghten Civic Competency and Engagement (Form 1)

Panelist	Proficient Round 1	Proficient Round 2	Advanced Round 1	Advanced Round 2
1	14.35	13.10	22.65	23.15
2	9.95	10.15	20.30	20.65
3	9.10	9.10	23.75	23.75
4	9.15	10.15	22.45	22.80
5	9.95	9.95	19.95	19.95
6	9.65	11.30	22.15	22.80
7	13.25	13.15	24.30	24.25
8	18.90	17.65	28.00	27.55
9	15.10	15.10	25.00	25.00
10	14.30	14.30	23.15	23.15
11	12.30	12.10	24.70	24.75
12	10.20	10.30	26.00	26.00
Mean	12.18	12.20	23.53	23.65
Minimum	9.10	9.10	19.95	19.95
Maximum	18.90	17.65	28.00	27.55
SD	3.06	2.55	2.29	2.11
Standard error of judgment	0.88	0.74	0.66	0.61

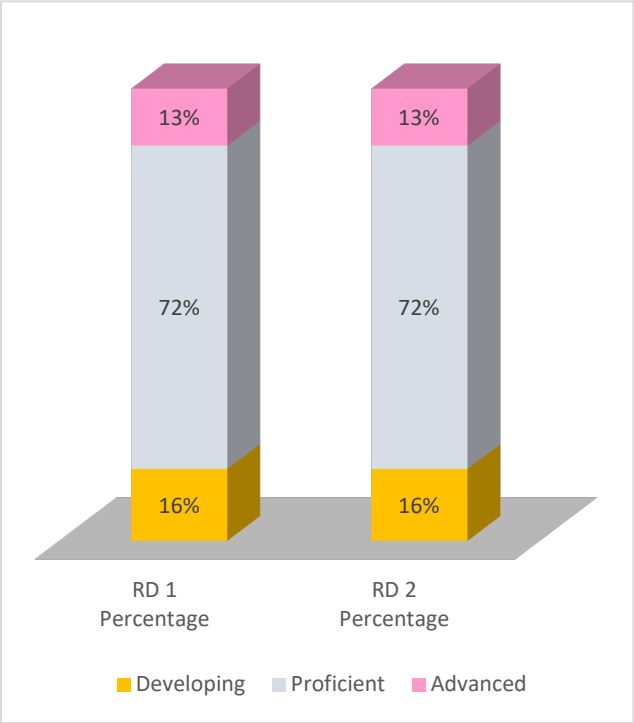

Figure F1. Impact data per round: HEIghten Civic Competency and Engagement (Form 1). *N* (seniors) = 151.

Table F2. HEIghten Civic Competency and Engagement (Form 2)

Panelist	Proficient Round 1	Proficient Round 2	Advanced Round 1	Advanced Round 2	Advanced, Round 2, outlier removed
1	13.65	13.40	23.70	23.60	23.60
2	8.15	9.45	17.15	18.90	–
3	6.60	7.35	21.70	22.20	22.20
4	9.10	10.15	22.45	22.85	22.85
5	8.60	8.60	20.35	20.35	20.35
6	8.50	9.50	22.35	22.80	22.80
7	12.85	12.75	23.80	23.60	23.60
8	17.45	16.75	25.60	25.45	25.45
9	11.90	12.15	22.05	22.05	22.05
10	14.85	14.75	21.55	21.55	21.55
11	13.50	12.75	26.20	25.70	25.70
12	8.45	8.90	24.15	24.05	24.05
Mean	11.13	11.38	22.59	22.76	23.11
Minimum	6.60	7.35	17.15	18.90	20.35
Maximum	17.45	16.75	26.20	25.70	25.70
SD	3.35	2.82	2.41	1.95	1.60
Standard error of judgment	0.97	0.81	0.70	0.56	0.48

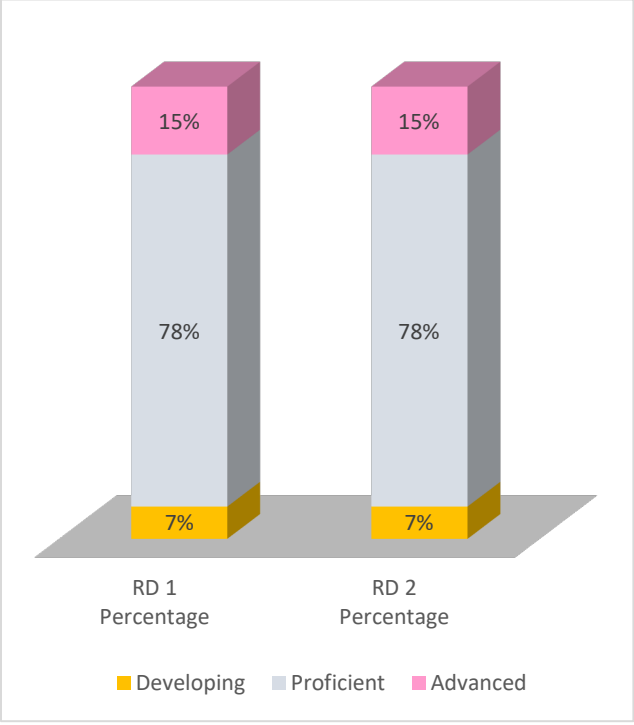

Figure F2. Impact data per round: HEIghten Civic Competency and Engagement (Form 2). *N* (seniors) = 140.

Table F3. HEIghten Intercultural Competency and Diversity (Form 1)

Panelist	Proficient Round 1	Proficient Round 2	Advanced Round 1	Advanced Round 2
1	16.50	16.40	29.70	29.55
2	17.70	16.90	31.75	31.90
3	10.55	10.85	30.95	31.05
4	17.55	17.25	32.95	32.85
5	18.80	18.80	31.35	31.35
6	13.80	14.25	30.95	31.25
7	18.70	18.50	33.25	33.15
8	20.40	20.20	33.40	33.25
9	15.20	15.40	31.20	31.20
10	19.15	18.95	31.70	31.70
11	13.95	18.55	31.65	34.55
12	14.40	14.95	34.50	34.75
Mean	16.39	16.75	31.95	32.21
Minimum	10.55	10.85	29.70	29.55
Maximum	20.40	20.20	34.50	34.75
SD	2.86	2.59	1.33	1.53
Standard error of judgment	0.83	0.75	0.38	0.44

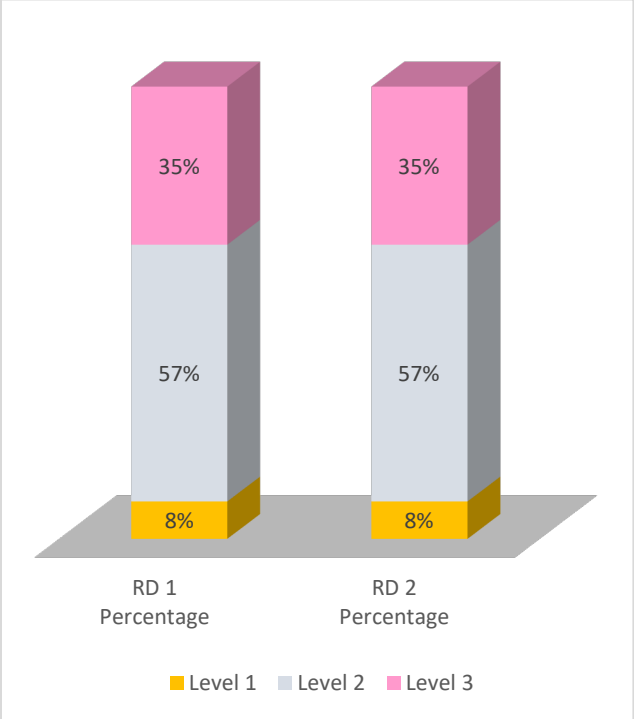

Figure F3. Impact data per round: HEIghten Intercultural Competency and Diversity (Form 1). *N* (seniors) = 169.

Table F4. HEIghten Intercultural Competency and Diversity (Form 2)

Panelist	Proficient Round 1	Proficient Round 2	Advanced Round 1	Advanced Round 2	Advanced, Round 2, outlier removed
1	16.60	15.65	28.90	28.10	–
2	17.40	16.80	31.30	31.30	31.30
3	9.75	10.15	31.60	31.70	31.70
4	15.55	15.75	32.55	32.55	32.55
5	18.05	18.05	31.10	31.10	31.10
6	13.00	13.45	30.25	30.65	30.65
7	18.60	18.60	33.35	33.35	33.35
8	18.70	18.50	32.95	32.80	32.80
9	12.00	12.10	31.30	31.30	31.30
10	20.05	19.90	32.35	32.35	32.35
11	12.55	16.50	32.50	34.60	34.60
12	12.10	12.45	33.50	33.60	33.60
Mean	15.36	15.66	31.80	31.95	32.30
Minimum	9.75	10.15	28.90	28.10	30.65
Maximum	20.05	19.90	33.50	34.60	34.60
SD	3.35	3.02	1.34	1.68	1.22
Standard error of judgment	0.97	0.87	0.39	0.49	0.37

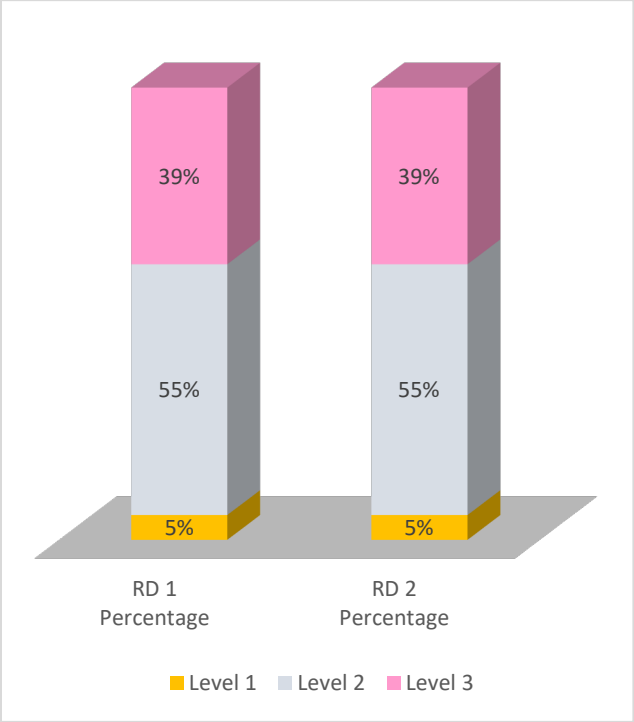

Figure F4. Impact data per round: HEIghten Intercultural Competency and Diversity (Form 2). *N* (seniors) = 182.

Appendix G. Final Evaluations

Table G1. Overall Evaluation

Statement	Number strongly agree	Number agree	Number disagree	Number strongly disagree
I understood the purpose of this study.	8	4	0	0
The instructions and explanations provided by the facilitators were clear.	10	2	0	0
The training in the standard-setting method was adequate to give me the information I needed to complete my assignment.	7	3	2[a]	0
The explanation of how the recommended score is computed was clear.	8	4	0	0
The opportunity for feedback and discussion between rounds was helpful.	6	5	1	0
The process of making the standard-setting judgments was easy to follow.	7	4	1	0
I understood how to use the survey software.	12	0	0	0

[a]Following the training and practice, all of the panelists indicated that they understood the training and were ready to proceed in making their standard-setting judgments. One of the judgments from one of these panelists was an outlier.

Table G2. Factors That Influence Judgments

How influential was each of the following factors in guiding your standard-setting judgments?	Number very influential	Number somewhat influential	Number not influential
The definition of the borderline student	9	3	0
The between-round discussions	5	7	0
The knowledge required to answer each question	9	3	0
The recommended scores of other panel members	2	8	2
My own professional experience	10	2	0

Table G3. Support of Final Recommended Levels

Do you believe that the final recommended levels are too low, about right, or too high?	Number too low	Number about right	Number too high
Civic Competency and Engagement, proficient	2	9	0
Civic Competency and Engagement, advanced	0	10	1
Intercultural Competency and Diversity, proficient	3	8	0
Intercultural Competency and Diversity, advanced	3	6	2

Note. One panelist left the study early and did not see the final results.

Table G4. Support of Final Recommendations

Do you support the final recommendations of the panel?	Number yes	Number no
Civic Competency and Engagement, proficient	11	0
Civic Competency and Engagement, advanced	11	0
Intercultural Competency and Diversity, proficient	10	1
Intercultural Competency and Diversity, advanced[a]	10	0

Note. One panelist left the study early and did not see the final results.
[a]A panelist expressed uncertainty and did not respond to this question.

Notes

[1]The two forms used for each assessment did not share common items. The two test forms are spiraled, and an equivalent groups equating design is used (U. Ali, personal communication, July 27, 2018).

[2]The Likert items for the Civic Attitudes dimension on each assessment are on a 4-point scale ranging from 1 (*strongly disagree*) to 4 (*strongly agree*).

[3]The selected-response items for Civic Participation include multiple-selection multiple-choice items about the context or location of the civic participation (e.g., on campus, at the national level), single-selection multiple-choice items about the number of hours volunteering, and Likert items on a 4-point scale ranging from 1 (*never*) to 4 (*daily*).

[4]The Likert items for the Approach dimension on each assessment are on a 4-point scale ranging from 1 (*strongly disagree*) to 2 (*strongly agree*).

[5]During the standard-setting study, the performance levels for the Analyze and Act dimension of the ICD assessment were titled Levels 1, 2, and 3. The decision to rename the levels developing, proficient, and advanced was made at a later time.

[6]This is true even if the test taker scores at a level equal to chance.

[7]The assessment specialists created PLDs for the entire assessment. Only the Civic Competency section is included in this report because it is relevant to the standard-setting study.

[8]The assessment specialists created PLDs for the entire assessment. Only the Analyze and Act section is included in this report because it is relevant to the standard-setting study.