Measuring the Power of Learning.®

Research Memorandum
ETS RM–17-06

Providing Threshold Score Recommendations for the First Three Tests of the *HElghten*® Outcomes Assessment Suite: A Standard-Setting Study

Wanda D. Swiggett

December 2017

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 First Three Tests of the *HEIghten*® Outcomes Assessment Suite: A Standard-Setting Study

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

December 2017

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

Suggested citation: 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.

Action Editor: Donald Powers

Reviewers: Patricia Baron and Katrina Roohr

Abstract

The *HEIghten*® outcomes assessment suite is made up of multiple assessments designed to assess general student learning outcomes for students exiting college. These assessments are modular and are designed to inform students as well as institutions of student progress on each learning outcome. A standard-setting study was conducted for the first 3 HEIghten assessments—critical thinking, quantitative literacy, and written communication. A panel of 11 college professors who teach undergraduate students participated in the study. These educators teach courses in which the general education skills assessed in the 3 HEIghten assessments can be attained. For each of the 3 assessments, the panel recommended 2 threshold scores, which mark the beginning of the *proficient* and *advanced* performance categories. Scores below the proficient threshold score classify students as developing. The panel provided recommendations for 2 forms of each assessment. The panelists made 2 rounds of item-level judgments, following the modified Angoff and extended Angoff methods, for the selected- and constructed-response items, respectively. Between each round, they discussed panelists' judgments and impact data based on the operational administration of the assessments. Evaluations were administered after each training and at the conclusion of the study. The panelists indicated their training in standard setting was clear, that the process was easy to follow, and that they understood the purpose of the study. Additionally, the panelists indicated that they supported all the final recommended scores.

Key words: standard setting, Angoff, *HEIghten*®, threshold scores, performance level descriptors, borderline students, impact data

Acknowledgments

The author would like to thank Kri Burkander, J. Wyman Brantley, Lois E. Frankel, Stephen Peterson, Lily Chiu, Usama Ali, Lauren Bauser, Heather Walters, Craig Stief, Lydia Liu, and Richard J. Tannenbaum. Their hard work was instrumental in preparing for and completing the standard-setting study.

Table of Contents

The *HEIghten*® outcomes assessment suite is made up of multiple assessments designed to assess general education student learning outcomes. The author designed and conducted the standard-setting study on May 11–13, 2016, on two forms of the first three HEIghten assessments—critical thinking, quantitative literacy, and written communication—to establish threshold scores, which mark the beginning of the *proficient* and *advanced* performance categories. Scores below the proficient threshold score classify students as *developing*. The terms *student* and *test taker* will be used interchangeably throughout this report.

The HEIghten program considered those recommended scores and other sources of information when setting the final threshold scores for these assessments. There were no correct decisions; the appropriateness of any adjustment—toward higher or lower scores—may be evaluated only in terms of its meeting the test's development and use (see Geisinger & McCormick, 2010). This report documents the standard-setting process and results.

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], 2016). These student learning outcomes are skills that students develop throughout their undergraduate years that can be transferred in real-world situations, including students' future careers. These assessments help provide institutions with actionable data that can be used for a variety of purposes. For example, institutions can provide evidence of their students' skills, compare their students' scores against those of other institutions, and use the data to make decisions that can improve all students' learning. In addition to the institutions' uses, students have opportunity to earn a microcredential documenting the level of skill demonstrated with these assessments (ETS, 2016).

Assessments in Phase 1

The first phase of the suite consists of three assessments—critical thinking, quantitative literacy, and written communication. Those assessments launched operationally in 2015. Two additional assessments—one for civic competency and engagement and another for intercultural competency and diversity are currently in development. HEIghten is designed to be modular and easy to use. Institutions can select the assessments, or combination of assessments, that meet their needs. Each assessment is designed to have an administration time of 45 minutes, which

allows it to be administered during a class period. The assessments can also be administered online via remote proctoring (ETS, 2016). The first phase of the suite was considered during the standard-setting study. For each of the three assessments, two forms were used.

Critical thinking. The HEIghten Critical Thinking assessment contains 26 selected-response items measuring two central aspects of critical thinking—analytical and synthetic skills. The analytical skills include analyzing an argument structure, evaluating an argument structure, and evaluating evidence and its use. The synthetic skills include developing valid or sound arguments and understanding the implications or consequences of information and argumentation (ETS, 2016; Liu, Frankel, & Roohr, 2014).

Quantitative literacy. The HEIghten Quantitative Literacy assessment contains 25 selected-response items measuring problem-solving skills in major mathematical content areas. The problem-solving skills include interpreting information; strategically evaluating, inferring, and reasoning; capturing relationships between variables; manipulating expressions and computing quantities; and communicating mathematical ideas. These skills are measured within four major mathematical content areas: (a) number and operations, (b) algebra, (c) geometry and measurement, and (d) probability and statistics (ETS, 2016; Roohr, Graf, & Liu, 2014).

Written communication. The HEIghten Written Communication assessment includes 25 selected-response items and one constructed-response item. This assessment evaluates four key dimensions of written communication. The first dimension, knowledge of social and rhetorical situations, includes audience awareness and writing for particular purposes, tasks, or contexts. The second dimension, knowledge of conceptual strategies, includes content development and organization as well as use and citation of sources and textual evidence. The third dimension, knowledge of language use and conventions, includes grammar usage, mechanics, and syntax, as well as word choice, tone, voice, and style. The fourth dimension is knowledge of the writing process, which includes drafting and revising (ETS, 2016; Sparks, Song, Brantley, & Liu, 2014).

Performance Level Descriptors

The performance level descriptors (PLDs) describe the knowledge, skills, and abilities of students who can be categorized into the developing, proficient, and advanced performance levels. The PLDs are a crucial part of the standard-setting study (Cizek, 2012; Cizek & Bunch, 2007; Perie, 2008). They 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 is part of the score-reporting documentation and has been published on the public website of the HEIghten outcomes assessment suite (www.ets.org/heighten) so that test takers and institutions can understand the score reporting (see ETS, 2016). The ETS assessment specialists for each content area, along with a small group of college faculty who served as outside advisors, developed the PLDs.

To develop the PLDs for the critical thinking, quantitative literacy, and written communication assessments, ETS assessment specialists formed a committee of educators with experience in teaching, at the high school and college levels, the content and skills that are measured on the assessments. The committee reviewed the content specifications and representative questions from the assessments. It then met to discuss the assessment and to develop consensus on (a) the level of proficiency needed to correctly respond to a given representative item and (b) a description of the features of each item that contribute to the proficiency level. Based on these discussions, ETS assessment specialists drafted the PLDs, which were then further refined through reviews by ETS assessment specialists and researchers.

During the development of the PLDs, it was determined that the proficient and advanced levels would be described in terms of the capabilities a typical student at those levels can demonstrate. Because developing was the lowest possible classification, the committee decided that descriptions of typical students in that category would be expressed as limitations such as a student might exhibit. The completed PLDs are listed in Appendix A.

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. The minimum scores are described as threshold scores because they differentiate the minimum score required to breach the threshold of the 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 be 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 the 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. It provides support for the claims that institutions can make about their students as well as the meaning of the microcredentials (i.e., badges) that can be earned (Papageorgiou & Tannenbaum, 2016).

Methodology

For this standard setting, an expert panel of judges followed a standard-setting design based on the modified Angoff (Brandon, 2004; Hambleton & Pitoniak, 2006; Plake & Cizek, 2012) and extended Angoff methods (Cizek & Bunch, 2007; Hambleton & Plake, 1995); panelists made two rounds of judgments with feedback between rounds (Reckase & Chen, 2012). The feedback prompted discussion from the panelists as they made the second round of judgments.

As described in the *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 time (e.g., Plake & Cizek, 2012; Tannenbaum, 2011; Tannenbaum & Kannan, 2015).

In addition to documenting the procedures of the standard-setting methods, the methods were assessed through 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 G, Technical Note 1). 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 requires 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 to be specific to particular college majors, an interdisciplinary faculty panel was recruited to provide judgments for all three assessments. Specifically, faculty with background expertise in subjects such as English composition or literature, world and U.S. history, social sciences, calculus, and statistics were contacted to participate in the study. In addition, because HEIghten is designed to be used at both 2- and 4-year institutions, faculty from both types of institutions were selected. 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 test and then developed the definition of the borderline students for the proficient and advanced performance levels. After training and practice, the panel completed two rounds of judgments with data between the rounds indicating 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[1] placed into each category based on the panel's recommendations (Margolis & Clauser, 2014). At the conclusion of the study, the panelists were asked to complete a final evaluation of the standard-setting process. An agenda describing the process is included in Appendix B.

Although a single standard-setting study was conducted for all three assessments, separate scores were recommended for each assessment. Additionally, two forms of each assessment were used during the standard-setting study. Separate scores were recommended for each form.

Premeeting information. Approximately 2 weeks before the panelists arrived for the standard-setting study, they were provided with premeeting materials. They were given descriptions of the HEIghten outcomes assessment suite as well as the content specifications for each assessment (along with links to the HEIghten website, if they were interested in learning

more). Additionally, they were provided with the performance level descriptors for each assessment. The panelists were asked to consider the PLDs and complete a small assignment designed to help them begin thinking about the borderline students. They were asked to write notes on the knowledge and skills students at the beginning of the proficient and advanced ranges would 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 took each assessment independently to understand what was being measured and the relative difficulty of the items, and to get a sense of the test takers' experience. They started by reviewing the critical thinking assessment. They reviewed each of the two forms independently, followed by a group discussion focused on the knowledge and skills measured and any differences they noticed between the two forms. The discussion then progressed to the student learning outcomes college seniors need (specific to critical thinking skills) when they are ready to move on to the next phase of their lives (e.g., advanced education, career, productive citizenship). At the conclusion of these discussions, the panelists repeated the test-familiarization process with the quantitative literacy assessment, concluding with the written communication assessment.

Borderline student definitions. The discussions about the assessments provided an opportunity for the panelists to consider the knowledge and skills measured and how test takers in the different performance levels would vary within those skills. That task was an important precursor to creating the borderline student definitions for the proficient and advanced levels for each assessment.

To begin, the standard-setting researcher explained to the panel members that each PLD describes every student in that performance level, from the least to the most able. By contrast, the borderline student is the student who marks the entry point (the borderline) into each performance level (proficient or advanced). It is the operational definition of each threshold score. When panelists make their standard-setting judgments, they consider the knowledge and skills each borderline student has and determine what score the borderline student would earn on the assessment. That score, by definition, must be the threshold score into that performance level.

The panelists worked together to first define the borderline proficient student. They worked in small groups, reviewing their notes from the premeeting assignment, and created a

draft description of the borderline student for each assessment. Then they worked as a large group to consolidate each small group's draft definitions into one consensus document. After they completed the definition for a proficient student, the entire panel worked together to create the definition for the borderline advanced student. The completed definitions are included in Appendix C.

Standard-setting judgments. Prior to making standard-setting judgments, panelists received training. 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 likelihood (probability or chance) 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, the less likely that the borderline student would answer the item correctly. The higher the value, the more likely 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.

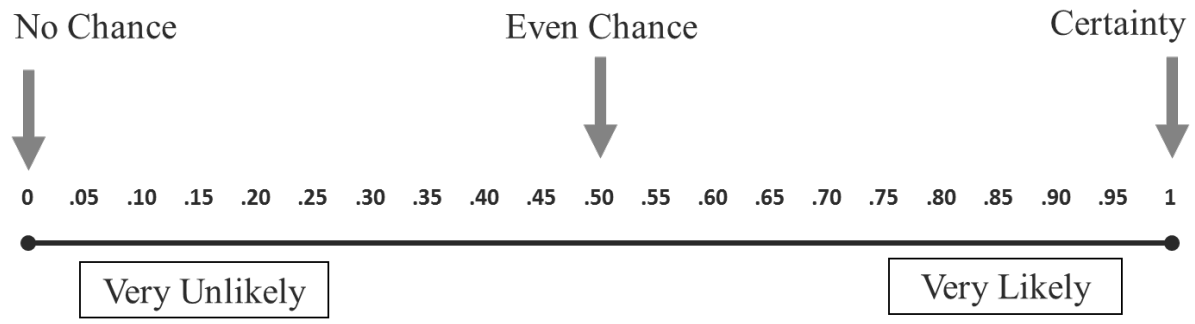

Figure 1. Modified Angoff judgments scale.

Panelists were asked to first review the description of the borderline student definition, beginning with proficient, and then to review the item. They were asked to decide 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 decision:

- Items in the 0 to .30 range were those the borderline judgment would have a low chance of answering correctly.

- Items in the .35 to .65 range were those the borderline judgment would have a moderate chance of answering correctly.
- Items in the .70 to 1 range were those that the borderline judgment would have a high chance of answering correctly.

Next, panelists decided how to refine their 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 to .65 range. The second decision for the panelist was to judge if the likelihood of answering it correctly was .35, .40, .45, .50, .55, .60, or .65 and indicate it on their rating form. They were reminded not to consider factors unrelated to the student's knowledge and skills.

After they completed their borderline proficient judgment for an item, they made their judgment for the borderline advanced student for the same item. They were reminded that, because borderline advanced students have a higher level of knowledge and skills, their judgment would have to have a higher probability than their borderline proficient judgment. The panelists followed the same judgment process, using the probability scale, for making their borderline advanced judgments.

After the training, panelists made practice judgments on a subset of items, followed by a group discussion of those judgments and their rationales. After the discussion concluded, all panelists completed a posttraining evaluation to confirm that they had received adequate training and felt prepared to continue; the standard-setting process continued only if all panelists confirmed their readiness. The panel provided judgments for all the selected-response items, beginning with the critical thinking assessments. The panelists did not require additional training or practice when they were ready to make their judgments on the selected-response items of the quantitative literacy and written communication assessments.

After the panelists completed their first round of selected-response judgments, they received training on how to make their extended Angoff (Cizek & Bunch, 2007; Hambleton & Plake, 1995) standard-setting judgments for the constructed-response items on the written communication assessments. For this portion of the study, panelists were instructed to assign a score value that would most likely be earned by the borderline proficient student for each constructed-response item. Panelists were asked to first review the definition of the borderline

proficient student and then to review the constructed-response item and its rubric. During this review, each panelist independently considered the level of knowledge/skill required to respond to the constructed-response item and the features of a response that would earn a particular score, as defined by the rubric. Each panelist decided on the score most likely to be earned by the borderline proficient student. After they made their borderline proficient judgment, they made their borderline advanced judgment. Again, they were instructed to provide a judgment higher than the judgment they provided for the borderline proficient student.

After the training, panelists made a practice judgment on one form 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 constructed-response judgments on the remaining form.

Multiple rounds of judgments. After the first (independent) round of judgments, 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 and constructed-response items were summed. The average of each panelist's recommendation was then calculated as the panel's threshold score recommendation.

The panel's judgments were displayed for each item and by the number of judgments in the low, moderate, or high probability ranges. For the selected-response items, the number of judgments were highlighted to show when at least two thirds of the panelist's judgments were in the same probability range. The panel members reviewed a selection of items and discussed the rationales for their judgments. These discussions helped panelists maintain a shared understanding of the knowledge and skills of the borderline students and helped to clarify aspects of items that might not have been clear to all panelists during the Round 1 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.

In addition to the summaries of their judgments, panelists were presented with the impact data[2] based on the operational administration 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.[3]

When presented with the impact data, the panel was informed as to what percentage of students would fall into each of the three categories given the recommended threshold scores from that first round of judgments. Panelists discussed their reactions to this information. At the end of this discussion, panelists were instructed to make their Round 2 independent judgments, keeping in mind all the feedback data, impact data, and discussions. The panelists entered data only if they were making a judgment that was different from their first-round judgment.

After the second round of judgments, the panelists reviewed the whole-group feedback as well as the impact data based on their Round 2 judgments. They discussed their final recommendations before completing the final evaluation form.

Evaluation of the process. As previously described, the panelists were provided with standard-setting training and practice followed by evaluations for each type judgment (i.e., modified or extended Angoff). In addition to these evaluations, the panelists completed a final evaluation form at the conclusion of the study. This form included questions about the standard-setting training, the between-round discussions, and the panelists' opinions about the recommended threshold scores for each of the assessments.

Results

Standard-Setting Panel

A panel of 11 college professors from eight states agreed to participate in the study. These educators have expertise in the learning outcomes measured in the critical thinking, quantitative literacy, and written communication assessments. They teach courses that a variety of students are likely to take, such as first-year writing, introduction to political science, and introductory mathematics. Their experience teaching these variety of courses is essential because students, regardless of major, are expected to gain these learning outcomes during their undergraduate education. The experts on the panel also have experience educating students from their freshman through senior years. They have experience seeing how students display the skills assessed in the HEIghten assessments throughout their undergraduate education.

The panel represented diverse demographic and professional backgrounds. They taught in small and large colleges and universities, private institutions, liberal arts colleges, and religious colleges. Most had been serving as faculty for over 10 years in either public or private 4-year institutions in various regions of the United States. The participants taught a large variety of

courses (e.g., writing, literary criticism, calculus, scientific computing, physics, and research methods) to first-year through senior college students as well as graduate students.

Table 1 shows the demographic summary of the standard-setting panel. The names and affiliations of the panelists are listed in Appendix D.

Table 1. Panel Demographics

Demographic	N
Current position	
Administrator/department head	1
College faculty	9
College faculty and administrator	1
Race	
White	7
Black or African American	2
Asian or Asian American	1
Prefer not to answer	1
Gender	
Female	4
Male	7
Students taught	
Freshmen/1st year, sophomores, juniors, seniors	8
Freshmen/1st year, juniors, seniors	1
Freshmen/1st year, sophomores	1
Seniors	1

Note: Six panelists also taught graduate students.

Standard-Setting Judgments

Table 2 summarizes the standard-setting judgments (Round 2) of panelists. Panelist-level results for Rounds 1 and 2 are presented in Appendix E, as well as impact data[4] based on their judgments from each round. The table also includes the standard deviation of the mean and the SEJ. 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 both in the standard deviation and SEJ. This decrease—indicating convergence among the panelists' judgments—was observed for each form of each assessment, with two exceptions. The SEJ for the advanced level of the quantitative literacy assessment (Form 1) remained the same from

Round 1 to Round 2. The SEJ for the advanced level of the written communication assessment (Form 1) increased, but the recommended score remained the same from Round 1 to Round 2.

Table 2. Recommended Threshold Scores by Round

Assessment and form code	Proficient, Round 1	Proficient, Round 2	Advanced, Round 1	Advanced, Round 2
Critical Thinking 1	12.64 (0.61)	12.61 (0.58)	19.76 (0.44)	19.77 (0.43)
Critical Thinking 2	11.27 (0.85)	11.19 (0.72)	18.74 (0.65)	18.76 (0.61)
Quantitative Literacy 1	10.66 (0.76)	10.61 (0.71)	18.76 (0.77)	18.79 (0.77)
Quantitative Literacy 3	9.82 (0.72)	9.93 (0.66)	18.63 (0.77)	18.62 (0.73)
Written Communication 1	17.69 (0.70)	17.74 (0.65)	27.67 (0.48)	27.67 (0.52)
Written Communication 2	17.45 (0.85)	17.40 (0.78)	27.12 (0.56)	27.29 (0.55)

Note. All threshold score recommendations are listed as raw scores. Standard errors of judgment are in parentheses. A conditional scoring rule was applied after the standard-setting study concluded (see Appendix G, Technical Note 2).

The panel's recommended threshold scores are the Round 2 mean scores. Table 3 notes the estimated conditional standard errors of measurement (CSEM; see Appendix G, Technical Note 3) around the recommended passing scores. A standard error represents the uncertainty associated with a test score. The conditional standard errors of measurement are estimates.

Table 3. Conditional Standard Errors of Measurement

Assessment and form code	Proficient	Advanced
Critical Thinking 1	12.61 (2.60)	19.77 (2.19)
Critical Thinking 2	11.19 (2.59)	18.76 (2.31)
Quantitative Literacy 1	10.61 (2.53)	18.79 (2.18)
Quantitative Literacy 3	9.93 (2.50)	18.62 (2.18)
Written Communication 1	17.74 (2.64)	27.67 (2.21)
Written Communication 2	17.40 (2.64)	27.29 (2.21)

Note. All threshold score recommendations are listed as raw scores and are based on the panel's final (Round 2) judgment. Conditional standard errors of measurement are in parentheses.

Evaluations

The panelists were given an evaluation after they were provided with training, practice, and discussion for each of the two types of judgments (modified Angoff and extended Angoff). On both posttraining evaluations, all 11 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. For each assessment, the panelists were provided the final recommended threshold scores and 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. For both levels of the critical thinking assessment, all the panelists indicated that the final recommendations were about right. One panelist believed that the quantitative literacy recommendation was too low for the proficient level; the rest of the panel indicated that the recommendations for both levels were about right. One panelist indicated that the final recommendation for the advanced level of the written communication assessment was too high. The rest of the panel indicated that the final recommendations for both levels of were about right. The entire panel, however, supported the final recommendations for both levels of all three assessments. A summary of the final evaluation results is presented in Appendix F.

On the final evaluation form, all panelists strongly agreed or agreed that they understood the purpose of the study. All the panelists strongly agreed or agreed that the facilitator's instructions and explanations were clear. All panelists strongly agreed or agreed that they were prepared to make their standard-setting judgments. All panelists strongly agreed or agreed that the standard-setting process was easy to follow. All 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.

Summary and Discussion

The author 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 proficient and advanced levels of three assessments in the HEIghten outcomes assessment suite. 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) according to Kane (1994, 2001).

The makeup of the panel is an essential part of the standard-setting study. This panel was made up of 11 educators with diverse backgrounds and experiences, who would have experience with the content measured on the assessment and also with students who would take the assessment. Though larger panels have been shown to provide consistent results, the size of the panel for this study is considered acceptable (e.g., Plake, Impara, & Irwin, 2000; Tannenbaum & Kannan, 2015). The 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. 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, the less likely it is that other panels would recommend a significantly different threshold score. The SEJs following Rounds 1 and 2 were low and support the consistency of the panelists' judgments and recommended threshold scores. The CSEMs of each form of each test also support the consistency of the scores associated with those forms. The tests were developed to be of equivalent difficulty, and the independent judgments from the panelists provide evidence supporting the development of those forms.

The HEIghten program accepted the recommendations of the standard-setting panelists, although they included a conditional scoring rule for the written communication assessment. The

scoring team was concerned that a test taker could be placed into the proficient or advanced performance level without making a reasonable effort on both item types (selected- or constructed-response). This concern was also expressed by at least one standard-setting panelist after the conclusion of the study. As a result, the scoring team implemented a conditional scoring rule (see Appendix G, Technical Note 2) stating that a test taker must earn at least a 6 (out of 12) on the essay in order to be categorized into the proficient or advanced performance levels, regardless of the total score earned. The decision was made after reviewing the standard-setting data and the data from the operational administration of the assessments.

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. Although these tests are not considered to be high stakes, it is still important that professional judgment be applied to the use of the assessments. In addition to now having threshold scores for the first three HEIghten assessments, the PLDs that were developed are also in use. They provide a rich description of the performance levels for the college students and institutions interested in the assessments. They are also part of the important information provided on the score reports, and they support the use of the test (e.g., students' microcredentials, an institution's evidence about its students).

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.

Cizek, G. J. (Ed.). (2012). *Setting performance standards: Foundations, methods, and innovations.* New York, NY: Routledge.

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

Educational Testing Service. (2016). *Introducing the* HEIghten® *Outcomes Assessment Suite*. Retrieved from http://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*, 38–44.

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.

Hambleton, R. K., & Plake, B. S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education, 8*, 41–55.

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, The Netherlands: CITO.

Kane, M. T. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64*, 425–462. http://dx.doi.org/10.3102/00346543064003425

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.

Liu, O. L., Frankel, L., & Roohr, K. C. (2014). *Assessing critical thinking in higher education: Current state and directions for next-generation assessment* (Research Report No. RR-14-10). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1984). Standard errors of measurement at different ability levels. *Journal of Educational Measurement, 21,* 239–275.

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,* 15–22. http://dx.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*(4), 295–317.

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

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.

Roohr, K. C., Graf, E. A., & Liu, O. L., (2014). *Assessing quantitative literacy in higher education: An overview of existing research and assessments with recommendations for next-generation assessment* (Research Report No. RR-14-22). Princeton, NJ: Educational Testing Service.

Sparks, J. R., Song, Y., Brantley, W., & Liu, O. L. (2014). *Assessing written communication in higher education: Review and recommendations for next-generation assessment*. (Research Report No. RR-14-37). 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., & 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.

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. http://dx.doi.org/10.1080/10627197.2015.997619

List of Appendices

Appendix A. Performance Level Descriptors

HEIghten Critical Thinking Performance Level Descriptions

Advanced

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

- extrapolate implications from multiple pieces of information and argumentation.
- accurately recognize descriptions of the logic of complexly structured arguments.
- employ multistep reasoning to identify hidden assumptions.
- employ multistep reasoning to identify evidence that directly or indirectly supports or undermines a claim, or specify additional information needed in order to resolve a point.
- identify subtle appeals to emotion and revisions to an argument that would reduce such appeals.
- distinguish information that may be peripherally or generally relevant to assertions or arguments from information that is directly on point.
- employ multistep reasoning to distinguish causation from correlation and identify possible alternative causes or explanations.
- engage in reasoning that involves complex interactions among multiple claims, arguments, or pieces of information.
- identify abstract concepts or principles that are implicitly instantiated in an argument.
- identify the most accurate among competing descriptions of the logical relationships between assertions or arguments and supporting (or irrelevant or undermining) information, even when the required distinctions are subtle or complex.

Proficient

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

- make inferential connections between points whose relationship is not explicitly given.

- follow the logic of an argument whose structure is not fully explicit.
- identify implicit assumptions.
- identify evidence that directly or indirectly supports or undermines a claim or specify additional information needed in order to resolve a point.
- identify appeals to emotion and revisions to an argument that would reduce such appeals.
- distinguish information that is relevant to assertions or arguments from irrelevant information.
- distinguish causation from correlation and identify possible alternative causes or explanations.
- engage in reasoning that involves interactions among multiple claims, arguments, or pieces of information.
- identify abstract concepts or principles that are instantiated in an argument.
- identify the most accurate among competing descriptions of the logical relationships between assertions or arguments and supporting (or irrelevant or undermining) information.

Developing

A typical student at the developing level may sometimes

- make inferential connections between two explicitly related points.
- follow the logic of an explicitly structured argument.
- identify explicit assumptions.
- identify evidence that directly supports or undermines a claim.
- identify clear appeals to emotion.
- mistake evidence that is broadly related to a topic for evidence that is relevant to a specific assertion about the topic.

- have difficulty distinguishing causation from correlation or identifying alternative explanations.
- have difficulty understanding or evaluating interactions among multiple claims, arguments, or pieces of evidence.
- have difficulty reasoning about abstract concepts or principles.
- have difficulty identifying the most accurate among competing descriptions of the logical relationships between assertions or arguments and supporting (or irrelevant or undermining) information.

HEIghten Quantitative Literacy Proficiency Level Descriptors

Advanced

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

- parse long, complicated word problems and extract relevant information to develop an appropriate model.
- recall and apply standard definitions, formulas, or algorithms that are appropriate for a given problem.
- set up and solve a model in a real-world context with two or three variables.
- solve multistep problems.
- recall and use basic algebra to solve equations that model a problem, e.g., use variables appropriately, manipulate and simplify algebraic expressions.
- recall and use basic facts of Euclidean geometry to model and solve problems, e.g., know formulas for perimeter, area, and volume; parallel and perpendicular lines.
- compute and interpret percents and percent change.
- read and interpret a chart or graph and extract data needed solve a problem.
- solve problems using proportional reasoning.
- perform the four basic operations (addition, subtraction, multiplication, and division) with integers, decimals, and fractions.

- recognize when there is insufficient information provided to solve a problem.
- read and interpret relationships between quantities expressed in terms of equations, formulas, or data representations.
- identify correct mathematical terminology and notation for communicating results.

Proficient

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

- reason through a problem in a real context, understand relevant nuances of context, and translate to an equation to solve.
- correctly use solution strategy of "plugging in appropriate numbers" or using a relevant example.
- set up a model in a real-world context with two or three variables, but may have difficulty solving the model.
- solve two- to three-step problems.
- recall and use basic algebra to solve equations that model a problem, but may have difficulty with algebraic manipulation.
- recall and use basic facts of Euclidean geometry to model and solve problems, but may not recall all the necessary facts.
- compute and interpret percents and percent change, but may have difficulty with percents greater than 100 and negative percent change.
- read a chart or graph, but may have difficulty interpreting the data presented.
- perform the four basic operations (addition, subtraction, multiplication, and division) with integers and decimals, but not necessarily fractions.
- choose appropriate variables for data in a problem, e.g., "let J be the number of cartons of juice purchased."
- read and interpret relationships between quantities expressed in terms of equations, formulas, or data representations, but may have difficulty with multiple variables or complex data representations.

- identify mathematical terminology and notation for communicating results, but may use incorrect terminology or notation.

Developing

A typical student at the developing level may sometimes

- parse simple word problems, but may react to surface features rather than apply quantitative reasoning.
- reason through a single-step word problem and translate to an equation to solve, but may have difficulty with complicated equations or calculations with large numbers.
- recognize when algebraic techniques are required to solve a problem, but may not recall the specific facts or techniques needed.
- recognize when facts from Euclidean geometry are required to solve a problem, but may not recall the specific facts or techniques needed.
- read a chart or graph, but may have difficulty extracting the data required to solve a problem.
- perform the four basic operations (addition, subtraction, multiplication, and division) with integers but not necessarily with decimals or fractions.
- read and interpret relationships between quantities expressed in terms of simple equations, well-known formulas, or simple data representations, but may have difficulty with multiple variables, new formulas, or complicated data representations.
- identify that mathematical terminology and notation are needed to communicate results, but may use incorrect terminology or incomplete notation.

HEIghten Written Communication Performance Level Descriptors

Advanced

A typical student at the advanced level has demonstrated

- the ability to compose or revise texts to successfully meet demands of purpose, audience, context, and task.

- the ability to successfully adhere to genre conventions such as argument and exposition/ explanation in writing or revising texts.
- the ability to easily navigate source texts in different genres and rhetorical modes.
- the ability to successfully incorporate or recognize the use of appropriate information from multiple source texts representing different genres to support their ideas.
- the ability to accurately represent a source's meaning, effectively using summary, paraphrase, and quotation, and to use or recognize appropriate citations.
- the ability to fully develop ideas or recognize the development of ideas using compelling reasons, examples, and evidence.
- the ability to effectively present ideas or recognize the effective presentation of ideas in an organized, logical, and coherent sequence in order to make complex ideas clear and understandable.
- the ability to effectively compose or recognize text that conveys meaning clearly by using engaging word choice, sentence variety, tone, voice, and style; what is appropriate will be determined by the context, purpose, and genre of writing.
- the ability to successfully compose or revise text to be free of all but minor errors in grammar, usage, mechanics, syntax, and spelling.
- mastery of the fundamental skills needed to produce fluent text.
- strategic knowledge of the writing process, including drafting, reviewing, revising, and editing.

Proficient

A typical student at the proficient level has demonstrated

- the ability, for familiar tasks and genres, to compose or revise texts to meet demands of purpose, audience, context, and task.
- the ability to adhere to genre conventions such as argument and exposition/explanation in writing or revising texts.
- the ability to navigate source texts in different genres and rhetorical modes.

- the ability to incorporate or recognize the use of appropriate information from source texts to support their ideas.
- the ability to represent a source's meaning with general accuracy, using summary, paraphrase, and quotation appropriately, and to use or recognize citations.
- the ability to develop ideas or recognize the development of ideas using sufficient reasons, examples, and evidence.
- the ability to present ideas or recognize the presentation of ideas in an organized, logical, and coherent sequence in order to make complex ideas clear and understandable.
- the ability to compose or recognize text that conveys meaning clearly by using appropriate word choice, sentence variety, tone, voice, and style; what is appropriate will be determined by the context, purpose, and genre of writing.
- the ability to compose or revise text to be generally free of errors in grammar, usage, mechanics, syntax, and spelling.
- command of the fundamental skills needed to produce fluent text.
- adequate knowledge of the writing process, including drafting, reviewing, revising, and editing.

Developing

A typical student at the developing level may

- have difficulty meeting demands of purpose, audience, context, and task, even for familiar tasks and genres.
- have difficulty adhering to genre conventions such as argument and exposition/explanation in writing or revising texts.
- not be able to navigate source texts in different genres and rhetorical modes.
- not consistently incorporate or recognize the use of appropriate information from source texts to support ideas.

- not be able to represent a source's meaning with general accuracy, using summary, paraphrase, and quotation appropriately, and may have trouble with citations.
- have difficulty developing ideas or recognizing the development of ideas using valid reasons and appropriate examples and evidence.
- struggle to present ideas or recognize the presentation of ideas in an organized, logical, and coherent sequence in order to make complex ideas clear and understandable.
- have difficulty composing or recognizing text that conveys meaning clearly by using appropriate word choice, sentence variety, tone, voice, and style; may struggle to know what is appropriate as determined by the context, purpose, and genre of writing.
- have difficulty composing or revising text to be generally free of errors in grammar, usage, mechanics, syntax, and spelling.
- demonstrate limited command of the fundamental skills needed to produce fluent text.
- lack sufficient knowledge of the writing process, including drafting, reviewing, revising, and editing.

Appendix B. Agenda

HEIghten Outcomes Assessment Suite
Standard Setting
Agenda
May 11–13, 2016

Day 1

8:00 AM Welcome and introductions

Overview of standard setting

Overview of the assessments

Break

Review Form A of quantitative literacy assessment and self-score

Review Form B of quantitative literacy assessment and self-score

Discuss the content measured on the forms

Discuss the quantitative literacy performance level descriptions (PLDs)

Lunch

Review Form A of critical thinking assessment and self-score

Review Form B of critical thinking assessment and self-score

Discuss the content measured on the forms

Discuss the critical thinking PLDs

Break

Review Form A of written communication assessment and self-score

Review Form B of written communication assessment and self-score

Discuss the content measured on the forms

Discuss the written communication PLDs

5:00 PM Collect materials; end of Day 1

Day 2

8:00 AM Overview for Day 2

Define borderline students (BLs) in relation to what is measured on the assessments

* Define proficient borderline student

* Define advanced borderline student

Break

Round 1 standard-setting training and practice for selected-response judgments

Complete Round 1 judgments for critical thinking assessments

Lunch

Complete Round 1 judgments for critical thinking assessments (*continued*)

Break

Data presentation and discussions

Complete Round 2 judgments for critical thinking assessments

Complete Round 1 judgments for quantitative literacy assessments

5:00 PM Collect materials; end of Day 2

Day 3

8:00 AM Overview for Day 3

Data presentation and discussions for quantitative literacy assessments

Complete Round 2 judgments for quantitative literacy assessments

Break

Complete Round 1 judgments for written communication assessments (selected-response items)

Lunch

Round 1 standard-setting training and practice for constructed-response judgments

Complete Round 1 constructed-response judgments for written communication assessments (constructed-response items)

Break

Data presentation and discussions for written communication assessments

Complete Round 2 judgments for written communication assessments (all items)

Break

Final results and discussion

Complete final evaluations

5:00 PM Collect materials; end of Day 3

Appendix C. Borderline Student Definitions

Proficient Borderline Student Definitions

Critical Thinking

1. Shows some basic ability to distinguish the degree of relevancy and accuracy of competing arguments.
2. Shows some basic ability to follow and evaluate the logic of an argument, including more explicit use of inferences, assumptions, and appeals to emotion.
3. Shows some basic ability to use multiple types of information to make or support an argument.
4. Shows some basic ability to distinguish causation from correlation.

Quantitative Literacy

1. Read and extract basic quantitative data from sources, including charts or graphs.
2. Convert basic information given into correct mathematical representations (e.g., text into mathematical representations, proportion).
3. Solve simple two-step problems.
4. Choose appropriate models to represent real-world situations.
5. Identify and use basic algebraic representations.
6. Perform the four basic operations with integers and decimals.
7. Identify and use basic geometric representations (rectangles, triangles, cubes).

Written Communication

1. Some basic ability to develop ideas using appropriate reasons, examples, evidence, and source materials.
2. Some basic ability to organize ideas in logical, coherent sequences.
3. Some basic ability to read and compose texts with an awareness of purpose, audience, and context for familiar tasks and genres.
4. Some basic ability to compose texts using standard writing conventions.

5. Some basic knowledge of writing process, including drafting, reviewing, and editing.

Advanced Borderline Student Definitions

Critical Thinking

1. Mostly consistent ability to distinguish the most relevant and accurate elements of competing arguments from multiple types of information.
2. Mostly consistent ability to evaluate the implicit logic of an argument, including use of inferences, assumptions, abstractions, and appeals to emotion.
3. Mostly consistent ability to use multistep reasoning, including extrapolation, to make and support an argument.
4. Ability to distinguish causation from correlation and identify possible alternative causes or explanations.

Quantitative Literacy

1. Extract and interpret relevant quantitative data from sources, including charts or graphs.
2. Set up and use a model to solve problems.
3. Identify and use algebra to solve equations and two-step and simple three-step problems.
4. Perform the four basic operations with integers, decimals, and simple fractions.
5. Identify and use basic geometric representations (area, volume, perimeter, parallel, perpendicular).
6. Use correct mathematical terminology and notation.
7. Solve problems using proportional reasoning.

Written Communication

1. Mostly consistent ability to develop ideas using apt reasons, examples, evidence, and source materials.
2. Mostly consistent ability to organize ideas in logical, coherent sequences.

3. Mostly consistent ability to read and compose texts with an awareness of purpose, audience, and context for various tasks and genres.
4. Mostly consistent ability to compose texts using standard writing conventions.
5. Emerging command of writing process, including drafting, reviewing, revising, and editing.
6. Mostly consistent ability to accurately represent a source's meaning, using summary, paraphrase, and quotation, and to use or recognize appropriate citations.

Appendix D. Participating Panelists and Affiliation

Panelist	Affiliation
Kodwo Annan	Georgia Gwinnett College (GA)
Warren Carson	University of South Carolina Upstate (SC)
Jose D'Arruda	University of North Carolina Pembroke (NC)
Van Hartmann	Manhattanville College (NY)
Marie Hoepfl	Appalachian State University (NC)
Jonathan Lang	University of California, Berkeley (CA)
Erin McNelis	Western Carolina University (NC)
Christopher Nelson	University of North Dakota (ND)
Ann Pelelo	Clarke University (IA)
Richard Strugala	Middlesex College (NJ)
Lisa Townsley	University of Georgia (GA)

Note. Panelists provided permission for their names and affiliations to be listed in this report.

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

Table E1. HEIghten Critical Thinking (Form 1)

Panelist	Proficient, Round 1	Proficient, Round 2	Advanced, Round 1	Advanced, Round 2
1	12.70	12.70	16.75	16.75
2	15.45	15.15	22.05	22.10
3	11.35	11.75	18.95	19.40
4	11.70	11.50	20.55	20.55
5	12.20	12.20	21.40	21.10
6	9.45	9.60	20.10	20.15
7	16.45	16.45	19.30	19.30
8	13.05	13.05	20.10	20.10
9	10.80	11.00	20.15	20.45
10	14.00	13.45	19.80	19.40
11	11.85	11.85	18.20	18.20
Mean	12.64	12.61	19.76	19.77
Minimum	9.45	9.60	16.75	16.75
Maximum	16.45	16.45	22.05	22.10
Standard deviation	2.03	1.91	1.46	1.44
Standard error of judgment	0.61	0.58	0.44	0.43

Note. *N*(seniors) = 391.

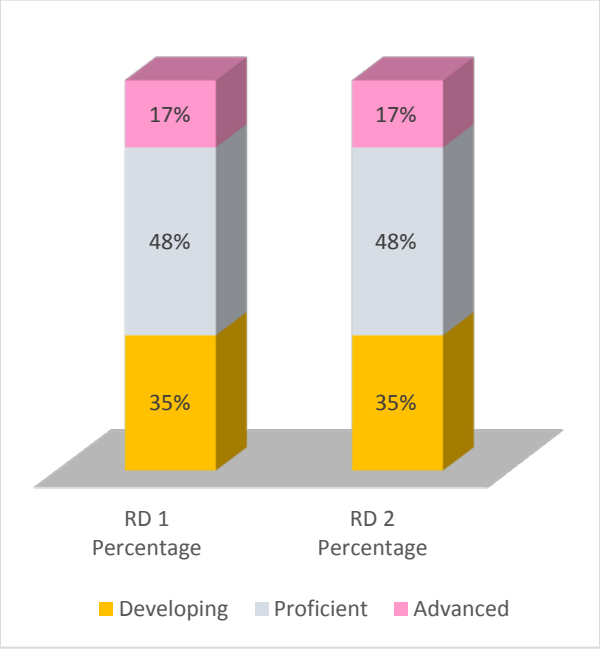

Figure E1. Impact data per round: HEIghten Critical Thinking (Form 1).

Table E2. HEIghten Critical Thinking (Form 2)

Panelist	Proficient, Round 1	Proficient, Round 2	Advanced, Round 1	Advanced, Round 2
1	9.85	10.05	15.45	15.45
2	14.15	13.80	20.75	20.90
3	7.50	7.85	14.65	15.35
4	9.35	9.35	19.50	19.50
5	13.35	13.35	22.30	22.30
6	8.00	8.40	18.70	18.70
7	16.85	15.35	19.35	18.80
8	11.95	11.95	19.20	19.20
9	9.45	9.40	19.10	19.00
10	11.80	11.80	19.20	19.20
11	11.70	11.75	17.95	17.95
Mean	11.27	11.19	18.74	18.76
Minimum	7.50	7.85	14.65	15.35
Maximum	16.85	15.35	22.30	22.30
Standard deviation	2.80	2.38	2.15	2.03
Standard error of judgment	0.85	0.72	0.65	0.61

Note. *N*(seniors) = 420.

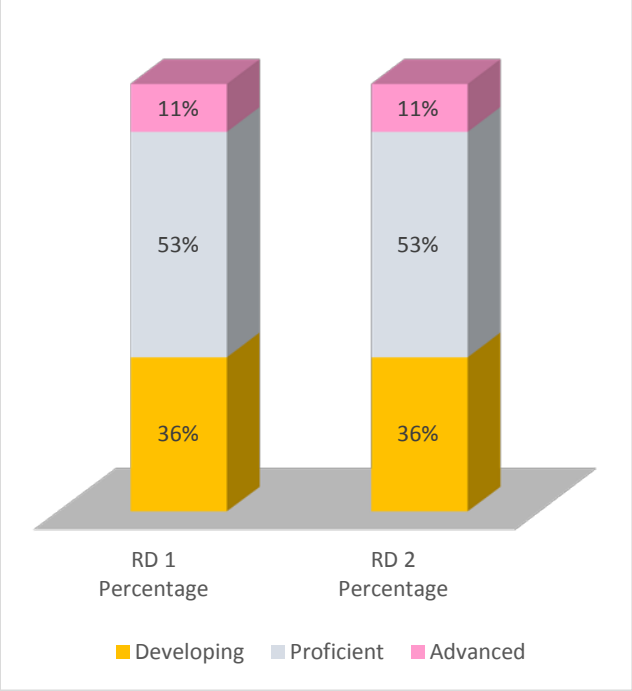

Figure E2. Impact data per round: HEIghten Critical Thinking (Form 2).

Table E3. HEIghten Quantitative Literacy (Form 1)

Panelist	Proficient, Round 1	Proficient, Round 2	Advanced, Round 1	Advanced, Round 2
1	8.85	8.85	14.80	14.80
2	6.20	6.60	15.40	15.40
3	9.75	9.55	18.45	18.35
4	12.80	12.70	21.40	21.40
5	12.45	12.45	20.45	20.45
6	10.05	10.05	19.50	19.70
7	9.05	9.05	14.85	15.00
8	15.25	14.75	21.20	21.20
9	8.65	8.65	20.65	20.65
10	12.25	12.25	20.40	20.40
11	11.95	11.80	19.30	19.30
Mean	10.66	10.61	18.76	18.79
Minimum	6.20	6.60	14.80	14.80
Maximum	15.25	14.75	21.40	21.40
Standard deviation	2.53	2.36	2.55	2.54
Standard error of judgment	0.76	0.71	0.77	0.77

Table E4. HEIghten Quantitative Literacy (Form 3)

Panelist	Proficient, Round 1	Proficient, Round 2	Advanced, Round 1	Advanced, Round 2
1	7.25	7.25	15.00	15.00
2	6.05	7.20	15.45	15.45
3	7.70	7.85	16.20	16.90
4	11.75	11.75	20.10	20.10
5	12.05	12.05	20.50	20.50
6	10.85	10.60	20.75	20.60
7	7.55	7.55	15.80	15.80
8	13.45	13.30	20.80	20.60
9	9.30	9.60	22.15	21.70
10	11.70	11.70	19.80	19.80
11	10.40	10.40	18.35	18.35
Mean	9.82	9.93	18.63	18.62
Minimum	6.05	7.20	15.00	15.00
Maximum	13.45	13.30	22.15	21.70
Standard deviation	2.40	2.18	2.57	2.42
Standard error of judgment	0.72	0.66	0.77	0.73

Table E5. HEIghten Written Communication (Form 1)

Panelist	Proficient, Round 1	Proficient, Round 2	Advanced, Round 1	Advanced, Round 2
1	17.20	17.20	24.70	24.70
2	21.10	20.70	28.80	28.80
3	16.75	17.05	28.10	28.10
4	16.05	16.05	27.85	27.85
5	15.75	15.75	27.85	28.85
6	17.60	17.70	28.00	27.95
7	20.55	20.55	27.10	27.10
8	20.70	20.70	30.60	30.60
9	13.90	14.45	26.60	25.60
10	18.75	18.75	29.00	29.00
11	16.25	16.25	25.80	25.80
Mean	17.69	17.74	27.67	27.67
Minimum	13.90	14.45	24.70	24.70
Maximum	21.10	20.70	30.60	30.60
Standard deviation	2.32	2.17	1.61	1.74
Standard error of judgment	0.70	0.65	0.48	0.52

Note. *N*(seniors) = 132.

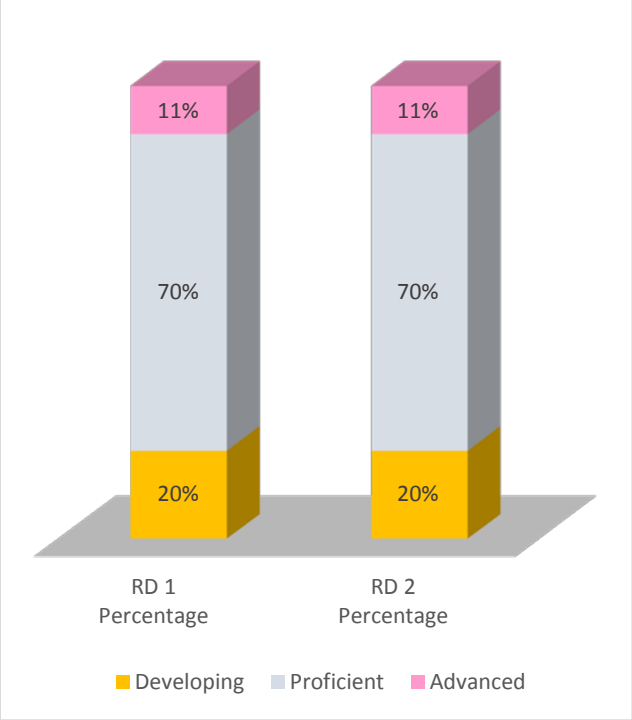

Figure E3. Impact data per round: HEIghten Written Communication (Form 1).

Table E6. HEIghten Written Communication (Form 2)

Panelist	Proficient, Round 1	Proficient, Round 2	Advanced, Round 1	Advanced, Round 2
1	17.05	17.05	23.70	23.70
2	20.90	19.80	27.55	28.55
3	14.20	14.40	27.05	27.20
4	16.70	16.70	27.30	27.30
5	14.90	14.90	27.90	27.90
6	16.80	16.70	29.30	29.30
7	21.00	21.00	27.95	27.95
8	19.85	19.80	29.75	29.75
9	13.75	14.30	24.85	25.50
10	21.15	21.15	27.90	27.90
11	15.65	15.65	25.10	25.10
Mean	17.45	17.40	27.12	27.29
Minimum	13.75	14.30	23.70	23.70
Maximum	21.15	21.15	29.75	29.75
Standard deviation	2.81	2.60	1.86	1.84
Standard error of judgment	0.85	0.78	0.56	0.55

Note. *N*(seniors) = 139.

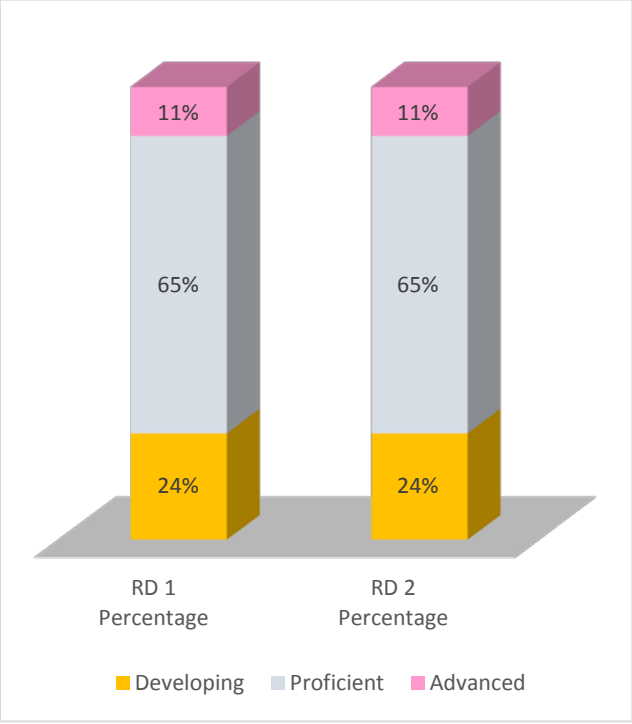

Figure E4. Impact data per round: HEIghten Written Communication (Form 2).

Appendix F. Final Evaluations

Table F1. Overall Evaluation

Statement	Strongly agree *N*	Agree *N*	Disagree *N*	Strongly disagree *N*
I understood the purpose of this study.	10	1	0	0
The instructions and explanations provided by the facilitators were clear.	10	1	0	0
The training in the standard-setting method was adequate to give me the information I needed to complete my assignment.	10	1	0	0
The explanation of how the recommended score is computed was clear.	8	3	0	0
The opportunity for feedback and discussion between rounds was helpful.	7	4	0	0
The process of making the standard-setting judgments was easy to follow.	10	1	0	0

Table F2. Factors That Influence Judgments

How influential was each of the following factors in guiding your standard-setting judgments?	Very influential *N*	Somewhat influential *N*	Not influential *N*
The definition of the borderline student definition	7	4	0
The between-round discussions	4	7	0
The knowledge/skills required to answer each assessment item	7	4	0
The recommended scores of other panel members	2	9	0
My own professional experience	9	2	0

Table F3. Support of Final Recommended Levels

Do you believe that the final recommended levels are too low, about right, or too high?	Too low *N*	About right *N*	Too high *N*
Critical Thinking, proficient	0	11	0
Critical Thinking, advanced	0	11	0
Quantitative Literacy, proficient	1	10	0
Quantitative Literacy, advanced	0	11	0
Written Communication, proficient	0	11	0
Written Communication, advanced	0	10	1

Table F4. Support of Final Recommendations

Do you support the final recommendations of the panel?	Yes *N*	No *N*
Critical Thinking, proficient	11	0
Critical Thinking, advanced	11	0
Quantitative Literacy, proficient	11	0
Quantitative Literacy, advanced	11	0
Written Communication, proficient	11	0
Written Communication, advanced	11	0

Appendix G. Technical Notes

1. Standard Error of Judgment (SEJ)

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. An 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).

2. Conditional Scoring Rule for the Written Communication Assessment

After the standard setting concluded, the HEIghten program implemented a conditional scoring rule regarding students' performance on the constructed-response item on the written communication assessment. Specifically, a student must earn at least a 6 (out of 12) on the essay to be considered proficient or advanced. Therefore, this scoring rule will be applied, in addition to the scores recommended by the standard-setting panel, to classify students into the performance levels.

3. Estimated Conditional Standard Error of Measurement (CSEM)

The estimated conditional standard error of measurement (CSEM) for a test consisting of both selected-response and constructed-response questions is

$$\text{CSEM} = \sqrt{(CSEM_{\text{SR}})^2 + (\text{SEM}_{\text{CR}})^2}$$

where CSEM_{SR} is computed from the study value (SV) of the recommended passing score and the number of selected-response questions (n) on the test (see Lord, 1984);

$$\text{CSEM}_{\text{SR}} = \sqrt{[SV\ (n - SV)] \,/\, (n - 1)}$$

and SEM_{CR} is computed as

$$\text{SEM}_{\text{CR}} = SD\sqrt{(1 - r)}$$

where the internal consistency reliability index, *r*, is set equal to .75 (a lower bound estimate) and the standard deviation (*SD*) is estimated as

$$SD = [(.95 * MAX) - MIN]/6$$

and MAX equals the maximum possible raw score for the constructed-responses questions and MIN equals the rounded value of (.05 * MAX).

Notes

[1] The impact data was based on the student performance available at the time of the study.

[2] For the quantitative literacy assessment, although the panelists viewed impact data upon request, they did not consider it due to the low sample size.

[3] A student is classified as unmotivated if (a) they fail to complete 75% or more of the test, or (b) the number of items answered in less than 10% of the mean item response time is equal to or greater than 20%.

[4] For the quantitative literacy assessment, the impact data are not included due to the low sample size.